Etude préliminaire en vue de la numérisation de la documentation scientifique de l'EPFL

Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES

Par:

Thierry USKE

Conseiller au travail de Bachelor :

Alexandre BODER, Chargé de cours HES

Genève, le 15 juillet 2011

Haute École de Gestion de Genève (HEG-GE)

Filière Information documentaire

Déclaration

Ce travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre de spécialiste en information documentaire. L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 15 juillet 2011

Thierry Uské

Remerciements

Je tiens à remercier particulièrement :

mon mandant, David Aymonin, pour son suivi et ses corrections au cours de l'élaboration du mémoire, sa disponibilité, ainsi que pour ses conseils et avis,

mon conseiller pédagogique, Alexandre Boder, pour les différentes indications et les pistes de réflexion qu'il m'a signalés,

mon juré, Jean-Marc Rod, pour avoir accepté d'évaluer mon mémoire,

mes professeurs HES, Jean-Marc Rod, Alexis Rivier et Jean-Daniel Zeller pour leurs précieux conseils et avis,

les différents collaborateurs de la bibliothèque de l'EPFL : Chantal Blanc, Alain Borel, Julie Chabloz Gachoud, Geneviève Freda Guéritault, Julien Junod, Simon Pasquier, Lionel Walter et les autres personnes que j'ai omis de citer pour leurs renseignements durant l'élaboration de mon travail,

Je remercie également les scientifiques des différents laboratoires de l'EPFL : Alain Nussbaumer, Professeur titulaire du Laboratoire de la construction métallique (ICOM), Olivier Burdet, Adjoint scientifique du Laboratoire de construction en béton (IBETON), Martin Schuler, Professeur titulaire de la Communauté d'études pour l'aménagement du territoire (CEAT), Henri-Pascal Mombelli, Associé de recherche au laboratoire des machines hydrauliques (LMH), qui m'ont présenté leur service et l'état des lieux de leur documentation.

Je remercie aussi les professionnels de la numérisation, Olivier Laffely, Responsable de l'atelier de numérisation de la Ville de Lausanne, ainsi que les collaborateurs de 4digitalbooks à Ecublens, pour leur disponibilité et leurs précieux conseils,

Mes remerciements vont également à Thomas Reynaud, Responsable du service de reprographie à l'EPFL, Patricia Plaza Gruber, Cheffe de projet GED à l'EPFL, Gregory Favre, Coordinateur technique d'Infoscience, pour leur accueil chaleureux et leur disponibilité.

Finalement je remercie Antoine et Danièle, pour leurs relectures et corrections.

Résumé

Cette étude préliminaire de numérisation vise à étudier si et comment il est possible de mettre en valeur les fonds et, indirectement, faire évoluer la gestion documentaire de plusieurs laboratoires scientifiques n'ayant pas migré totalement leurs collections à la bibliothèque de l'EPFL.

Mon travail consiste, dans un premier temps, à prendre contact avec un échantillon représentatif de laboratoires afin de réaliser un état des lieux de leur documentation et de la gestion de celle-ci.

Après analyse, des critères de choix et de priorisation dans le traitement des documents sont définis de concert avec les scientifiques et les bibliothécaires dans le cadre d'un projet de numérisation éventuel. Cette étape m'a permis également de prendre connaissance des besoins et attentes des scientifiques dans le domaine de la recherche documentaire.

Ensuite, mon étude s'intéresse au processus de numérisation dans sa globalité avant de se recentrer sur chaque entité étudiée. Quels documents méritent d'être scannés ? Qu'en est-il des droits d'auteur ? Peut-on lancer le projet en interne ou est-il préférable de s'orienter vers un prestataire externe ? Quels sont les coûts d'un tel projet ? L'outil Infoscience, l'archive institutionnelle de l'EPFL, est-il adapté au signalement et à la diffusion des documents numérisés ? Ces questions sont abordées au fil de mon étude tout en y apportant des éclaircissements et des éléments de réponse.

Une analyse critique de la situation est également menée en comparant les pratiques de l'EPFL avec d'autres institutions universitaires.

Ce rapport établit un constat actuel de la numérisation et aboutit à des recommandations et des axes d'amélioration dans le but de créer un guide de bonnes pratiques utile à la bibliothèque de l'EPFL et à d'autres laboratoires concernés par cette problématique.

Table des matières

Liste des Tableaux

Liste des Figures

Introduction

Présentation du mandat

Au fil de leur existence, les bibliothèques et laboratoires de l'EPFL (récemment réunies au sein d'une même entité dans le Rolex Learning Center) ont réuni une importante documentation sur les divers projets de recherche, d'enseignement et d'organisation administrative de l'EPFL (rapports, actes de congrès, travaux d'étudiants, etc.). Ces documents font partie de la mémoire et du capital scientifique de l'EPFL, cependant ils restent peu accessibles car disséminés dans plusieurs collections disparates et invisibles sur le web.

La bibliothèque de l'EPFL, récemment créée par la réunion au sein d'une même entité des dix plus grandes bibliothèques de l'EPFL, a formulé le postulat suivant : la numérisation de ces documents leur offrirait une meilleure accessibilité pour la communauté EPFL, les chercheurs du monde entier et le grand public et permettrait de ramener au jour des quantités d'information scientifiques de première qualité, tout en assurant une forme de conservation plus satisfaisante et de présentation cohérente.

Les objectifs du mandat

Les objectifs généraux de ce travail de diplôme consistent à améliorer la visibilité et l'accessibilité de la documentation des laboratoires de l'EPFL et de développer une expertise et un savoir-faire à la bibliothèque en matière d'analyse de fonds à numériser.

Plus spécifiquement, il s'agit dans un premier temps d'élaborer un état des lieux des collections importantes à numériser – à savoir établir une analyse sommaire de l'existant – et dans un deuxième temps, d'étudier les solutions techniques de numérisation afin de proposer un plan d'action pour la mise en ligne des documents à numériser.

Ces objectifs généraux se déclinent ainsi selon les objectifs spécifiques suivants :

- Elaborer un état des lieux des collections importantes à numériser
- Mener une revue des meilleures pratiques professionnelles dans le monde et en Suisse en matière de numérisation
- Etudier les solutions techniques de numérisation
- Proposer un plan d'action pour la mise en ligne des documents à numériser.

Les contraintes

Afin de répondre à ces objectifs, j'ai dû faire face à différentes contraintes liées aux organismes présents sur le campus. Plusieurs laboratoires gèrent leur documentation avec une certaine autonomie et des procédures qui leur sont propres. Par ailleurs, les catalogues de leurs centres de documentation ne sont parfois plus alimentés depuis plusieurs années maintenant et il est donc impossible d'obtenir un inventaire exhaustif des documents. Mon étude est donc partiellement basée sur des estimations lorsqu'il s'agit d'établir un état des lieux des collections.

En parallèle à cela, il n'était pas toujours aisé de planifier des rendez-vous avec les professionnels qui ont souvent d'autres priorités et peu de temps à consacrer à mon étude. Malheureusement plusieurs entretiens ont dû être annulés pour des raisons d'emploi du temps. En ce qui concerne les prestataires de numérisation, ces derniers n'étaient pas toujours en mesure de me présenter un devis ni une estimation des coûts sans avoir préalablement consulté les fonds à traiter.

De façon plus générale, il est à noter que la numérisation implique une attention particulière à la notion de droit d'auteur et sa variante le copyright. Nombre de documents sont soumis à une protection juridique. Il est donc nécessaire de s'entendre avec les auteurs et éditeurs avant de diffuser de manière globale les documents numérisés. Cela peut avoir des conséquences sur le mode de diffusion des documents, et réduire l'utilité d'un projet si les documents ne sont pas largement accessibles à la communauté scientifique. Les négociations avec les ayants droit (autorisation de diffusion) peuvent retarder considérablement l'avancement du projet.

Ma stratégie de travail a dû prendre en compte le fait que mon étude nécessite de nombreux entretiens auprès des différentes institutions concernées. Programmer ces rendez-vous n'est pas toujours chose aisée et cela a parfois prolongé les délais préalablement établis. De plus, au fil de l'avancement de mon travail, de nouveaux interlocuteurs se sont présentés ce qui a rendu le planning très serré.

Bibliothèque de l'EPFL

> *« L'EPFL est une université technologique qui accueille aujourd'hui près de 6.000 étudiants, dont plus de 1.000 doctorants. 3.000 chercheurs et scientifiques y développent des enseignements et des recherches en mathématiques, physique, chimie, management de la technologie, sciences de l'ingénieur, sciences de la vie, architecture. »*

(Source : http://infoscience.epfl.ch/record/125618/files/N1P2A2.pdf)

Ouverte au public le 22 février 2010, La bibliothèque de l'EPFL, constituant le cœur du Rolex Learning Center, est le fruit du regroupement des 10 plus grandes bibliothèques autrefois disséminées sur le site de l'EPFL. Conçu par le bureau d'architecture japonais SANAA, le RLC est le nouveau bâtiment phare du campus de l'EPFL ouvert 7 jours sur 7 de 7H à minuit. Hormis la bibliothèque, il abrite notamment les Presses Polytechniques Universitaires Romandes (PPUR), le Centre de Recherche et d'Appui pour la Formation et ses Technologies (CRAFT), le centre de carrières, une salle de conférence-spectacles de 400 places, trois restaurants, une librairie et une banque.

Les missions et les prestations de la bibliothèque

> *« Bibliothèque publique spécialisée dans les domaines d'étude et de recherche de l'EPFL, elle offre à chacun de ses utilisateurs un appui et des services utiles à la réussite de ses études, de ses recherches ou de son enseignement.*
>
> *« 900 Places de travail équipées d'un accès au réseau wifi, d'une prise de courant et d'une prise informatique. Photocopieuses et scanners. Dix salles de travail en groupe de différentes capacités. Plusieurs catalogues en ligne. »*
>
> (Source : http://www.unil.ch/bibliotheques-vaudoises/page61807.html)

En juin 2011, la bibliothèque emploie 45 professionnels (*2 ETP) et propose à ses lecteurs des collections papier et numérique : 500'000 documents (monographies + périodiques), 35'000 e-books, 11'000 périodiques électroniques, 600 abonnements papier, 5'000 thèses EPFL en ligne, 10'000 nouveaux documents par an, 70'000 publications EPFL (archive institutionnelle Infoscience).

Parmi ses services elle propose la numérisation d'articles de périodiques à la demande pour ses chercheurs.

Elle offre en libre service deux scanners à plat et des copieurs-scanners multifonction noir et blanc et couleur.

Méthodologie générale

Il s'agira tout d'abord d'élaborer un état des lieux des collections importantes à numériser, en définissant des critères de choix et de priorisation, en dialogue avec les bibliothécaires scientifiques et avec les détenteurs de ces collections dans les facultés.

Puis, de mener une revue des meilleures pratiques professionnelles dans le monde et en Suisse en matière de numérisation.

Cette étape sera suivie de l'étude des solutions techniques de numérisation et portera sur les capacités internes à l'EPFL (scanners de la bibliothèque, prestations de la

reprographie) et des prestataires existant dans le proche environnement à Lausanne et en Suisse romande : services, administrations, bibliothèques, sociétés spécialisées dans la numérisation. Leurs prestations, technologies, produits, coûts et processus seront analysés et comparés.

A l'issue de ces étapes de découverte des besoins et des solutions, nous proposerons un plan d'action (moyens nécessaires, choix techniques, délais, opérateurs, etc.) en vue de la mise en ligne de la documentation à numériser dans l'archive institutionnelle de l'EPFL ou sur une autre plateforme le cas échéant.

A terme, cette étude permettra de brosser un panorama des différents procédés de numérisation. Les différentes contraintes liées à un tel projet seront étudiées et des pistes de réflexion seront proposées. Les recherches seront menées en gardant pour objectif d'offrir une visibilité maximum au capital scientifique de l'EPFL.

La pluralité des centres de documentation et le temps limité octroyé à la rédaction de mon étude ne m'ont pas permis de procédé de manière exhaustive. En effet, je me suis orienté vers une démarche par échantillonnage des cas traités. Cela a été possible grâce aux connaissances du terrain des bibliothécaires.

Contacts avec le mandant

Suite au premier entretien avec mon mandant, il a été déterminé que David Aymonin superviserait mon mémoire.

Par la suite, j'ai pris contact de manière régulière avec M. Aymonin pour lui présenter l'avancement de mon travail et partager mes expériences sur le terrain.

J'ai eu également des contacts avec les nombreux bibliothécaires présents à la bibliothèque de l'EPFL ainsi que les responsables des différents laboratoires impliqués dans mon étude.

Les entrevues entre le mandant et moi-même se planifiaient en fonction des besoins, de manière sporadique. Durant ces instants, je présentais un compte-rendu de l'avancement de mon rapport et je profitais de cette occasion pour demander des éclaircissements et des conseils sur l'orientation de mon travail.

Organisation du travail et outils

Après avoir clarifié le mandat avec les parties prenantes, c'est en faisant une revue de la littérature sur le thème de la numérisation que m'on étude à réellement débuté. Je

me suis orienté tout d'abord vers des ouvrages généraux ainsi que des travaux de diplôme traitant de près ou de loin ma problématique. J'ai consulté les catalogues du réseau des bibliothèques de Suisse occidentale (RERO) et du réseau de bibliothèques et centres d'information en Suisse (NEBIS). Cette documentation m'a été utile pour établir un panorama de la numérisation en Suisse et dans le monde. C'est en consultant la bibliographie de ces ouvrages que j'ai découverts d'autres sources d'information spécialisées dans le domaine de la numérisation. Des recherches sur Internet ont permis de préciser certains points et de découvrir des comptes-rendus d'expérience de différentes institutions en relation directe avec ma problématique.

Afin de réunir les informations nécessaires à l'analyse et l'état des lieux des collections, j'ai consulté les ressources électroniques telles que les bases de données bibliographiques des laboratoires, leurs catalogues (si disponible) ainsi que les différentes bases de données spécialisées comme Inspec et Compendex (Engineering Village).

Voici une présentation des principales sources électroniques consultées :

- Les banques de données: Lexis Nexis, Dialog, Lisa, Lista, Engineering village (Inspec + Compendex + Referex).

Mots-clés: digitization, digital, metadata, metadata harvesting, data processing, OAI-PMH, Dublin Core, XML, guideline, library storage center, digital repository, retrieval system, automation, cost.

Je me suis rapidement rendu compte qu'il était nécessaire de se limiter à un choix restreint d'outils. J'ai sélectionné finalement les bases de données de l'Engineering village, c'est-à-dire Inspec, Compendex et Referex ainsi que Lista. Les références découvertes via ces outils m'ont été d'une aide précieuse pour appréhender le monde de la numérisation. Cependant, plusieurs occurrences traitaient d'aspects techniques avancés dans ce domaine. Il m'était donc difficile par moment de vulgariser ces informations afin de rendre accessible mon étude aux profanes.

Afin de structurer les occurrences récoltées, j'ai créé des dossiers thématiques pour le dépôt des documents dans Engineering village et Lista. Ceci m'a permis d'organiser l'information pour faciliter la consultation ultérieure.

- Les portails d'information: Infoscience, portail de l'ingénieur, digicoord, admin.ch
- Les sites web: IFLA, BBF, EPFL et institutions universitaires, etc.

- Les blogs, forums: bibliobsession, bibliothèques numériques, blogs d'institutions universitaires, etc.

En menant mes recherches, j'ai découvert de nombreux didacticiels et autres guides de bonnes pratiques dans le domaine de la numérisation. Des comptes rendus d'expérience m'ont permis de cibler plus aisément les multiples contraintes inhérentes à un tel projet.

Après avoir emmagasiné ces informations, j'ai pris contact avec les différents laboratoires afin d'entrer dans le vif du sujet. Dès lors, je me suis entretenu avec les responsables pour élaborer un état des lieux sommaire de leurs collections. En parallèle, j'ai effectué des visites dans les locaux de différents professionnels de traitement d'image afin de bénéficier de leur expérience en la matière. Pour préparer au mieux mes entretiens, j'ai élaboré une grille de questions permettant de « standardiser » la récolte d'information.

Les nombreuses discussions avec les bibliothécaires m'ont permis de préciser les points d'ombre entourant les laboratoires. Ils m'ont éclairé également sur l'historicité de certains fonds ainsi que sur le contexte pré Learning Center. Cela m'a été d'une grande aide lorsqu'il a fallu établir une sélection de documents suivant plusieurs critères de priorisation.

J'ai également bénéficié des discussions entre camarades traitant de près ou de loin le thème de la numérisation pour leurs propres travaux. Cela a apporté de nouvelles pistes à mon étude.

Bien entendu, j'ai développé des techniques d'enquêtes telles que les entretiens et autres visites afin de prendre connaissance des besoins et attentes de la communauté scientifique de l'EPFL.

Tous les comptes-rendus de mes entretiens et visites sont disponibles sur le Wiki de travail mis en place dans le cadre de ce projet : http://wiki.epfl.ch/etude-numerisation

Cet outil m'a permis d'organiser et de structurer mon travail ainsi que de garder une trace de mes recherches. Il peut s'apparenter à une sorte de carnet de bord. Cette centralisation de l'information a également facilité le suivi pour mon mandant.

La structure du Wiki est la suivante :

- Mandat
- Cahier des charges
- Bibliothèque de l'EPFL

- Etat des lieux des collections
- Site web & Infoscience
- Numérisation
- Méthodologie
- Agenda
- Bibliographie
- Glossaire

L'onglet « Mandat » présente les objectifs et la démarche globale de l'étude préliminaire de numérisation.

L'onglet « Cahier des charges » contient mon propre cahier des charges présentant les grandes lignes de mon étude.

L'onglet « Bibliothèque de l'EPFL » décrit l'historique et les missions de cette institution.

L'onglet « Etat des lieux des collections » présente les différents laboratoires entrant dans le cadre de mon étude ainsi que l'état des lieux de leur documentation.

L'onglet « Site web et Infoscience » présente dans les grandes lignes les sites Internet de l'EPFL et l'archive institutionnelle Infoscience.

L'onglet « Numérisation » décrit les principes et étapes de ce processus.

L'onglet « Méthodologie » illustre le fonctionnement de mes recherches.

L'onglet « Agenda » regroupe les dates de mes entrevues avec les différents acteurs du projet ainsi que les professionnels de la numérisation.

L'onglet « Bibliographie » recense les documents et les sources internet consultés pour le développement de mon étude.

L'onglet « Glossaire » apporte un éclaircissement sur la terminologie et les sigles utilisés dans mon rapport. Ce glossaire est repris en annexe au présent rapport[1].

[1] Glossaire : voir annexe 1

1. La numérisation

1.1 Historique

Les débuts de la numérisation coïncident avec les débuts de l'informatique (au 20e siècle). Les premiers pas sont très lents et l'accélération survient dans les années 90 avec l'invention du web et l'augmentation des performances des ordinateurs. La façon de rechercher l'information change et l'apparition de la génération Google met Internet en première ligne pour la diffusion de l'information. La pratique de la lecture à l'écran ainsi que la navigation hypertextuelle modifient en profondeur les paradigmes passés. De ce fait, la numérisation des données est un véritable enjeu pour la diffusion des connaissances et pour la démocratisation de la culture scientifique à l'échelle mondiale. Elle est entrée dans les pratiques courantes des professionnels et a un impact direct sur le rôle des bibliothèques et sur l'évolution de leurs missions.

En 2004, Google crée la surprise en lançant un projet de numérisation de plusieurs millions de livres intitulé Google Livres (ou Google Books) et qui vise une position dominante dans le commerce du livre numérique avec plus de 12 millions de volumes numérisés. Ce projet à créer des émules dans le monde et en Europe. Tout d'abord, c'est en 2006 que nait le projet Europeana. Cette bibliothèque numérique européenne vise à contrer ce qui est perçu par certains comme l'impérialisme documentaire de la firme de Mountain View.

> *« Concrètement, Europeana est une mise en commun des ressources (livres, matériel audiovisuel, photographies, documents d'archives, etc.) numériques des bibliothèques nationales des 27 États membres. Pour cela, les États s'engagent à numériser leurs contenus actuellement conservés de manière traditionnelle, les rendre accessibles sur le Web et assurer la conservation de ceux-ci sous forme numérique pour les générations futures. Le projet prévoit de faire appel, outre les bibliothèques nationales, aux bibliothèques européennes, aux services d'archivages et aux musées. »*

(Source : wikipedia.org)

Le projet Gutenberg, initié en 1971 par Michael Hart de l'Université de l'Illinois aux Etats-Unis, fait office de pionnier dans le domaine. Il recense des versions électroniques de livres physiquement existants faisant partie du domaine public.

La bibliothèque numérique de la Bibliothèque Nationale de France (BNF), nommée Gallica, propose en libre accès des livres numérisés, des revues, des photos et d'autres types de document depuis 1997.

C'est en 2007 que le Réseau des bibliothèques de Suisse occidentale (RERO) et la Bibliothèque nationale suisse (BNS) ont décidé de collaborer pour mettre à disposition une plateforme d'information sur les projets de numérisation suisses, qu'ils soient au stade d'une intention, en cours ou réalisés. Le but de ce projet est de pallier la difficulté de trouver les informations utiles dans ce domaine en Suisse.

Même si la structure fédéraliste de la Suisse et le primat des Cantons en matière culturelle, ne permettent pas la mise en place de projet de numérisation d'envergure nationale, les projets locaux ou issus d'une coopération entre institutions se développent. Ainsi, la Bibliothèque nationale suisse ne reste pas inactive dans ce domaine et soutient des projets d'intérêt national sous la forme de partenariats publics-privés. Par ailleurs, le projet e-lib.ch[2] suivi par la Conférence des bibliothèques universitaires suisses développe de nouvelles ressources numérisées en vue d'arriver à constituer la bibliothèque numérique suisse. La participation des institutions des EPF est assurée par le Conseil des EPF. Ce projet contient de nombreuses ressources numérisées :

- e-codices : Bibliothèque virtuelle de manuscrits suisses.
- e-rara.ch : Plate-forme en ligne pour des imprimés anciens numérisés dans des bibliothèques suisses.
- Retro-Seals : Plate-forme pour des périodiques numérisés.

Il comprend, de plus, un recensement des « best practices » à l'adresse suivante : http://www.digitalisierung.ethz.ch/

On notera également la participation de la Bibliothèque cantonale universitaire de Lausanne (BCU) au projet Google Livres avec plus de 100'000 titres en cours de numérisation.

1.2 Les apports de la numérisation

La banalisation du document numérique dans le domaine scientifique est un fait acquis désormais. On le remarque lorsque l'on analyse de plus près les bibliographies d'études scientifiques. Les références signalent désormais majoritairement des documents accessibles sous forme numérique ou des sites web au détriment de documents papier. Nous faisons face à un réel bouleversement documentaire. Avec le document numérique, on passe d'une logique de stock à une logique de flux et d'une gestion de document à une gestion de contenu.

[2] E-Lib. *E-Lib* (en ligne). 2011. http://www.e-lib.ch/info_f.html (consulté le 15.05.2011)

Chaque jour, le document numérique s'installe un peu plus dans nos activités. Souvent mis en avant, les enjeux des projets de numérisation sont multiples:

- l'amélioration des services rendus aux usagers
- la diffusion et la valorisation de corpus
- la préservation et la conservation des collections

Il est donc nécessaire de développer des contenus et des moyens d'accès répondant aux attentes du public et de proposer des nouveaux usages pour les documents numérisés en prenant compte des possibilités de coopération entre institutions.

La numérisation rend visible les ressources éloignées ou difficiles d'accès, elle permet de constituer des collections virtuelles avec des ressources éparpillées dans le monde. En outre, elle donne des possibilités de recherche très étendues comme la recherche booléenne, la recherche dans le texte, la recherche par champs etc. Un des attraits principaux de cette technologie est la possibilité de collaboration étendue entre les institutions. Le traitement des publications en ligne et la masse d'information que génère internet rend impossible le fait de tout collecter pour une seule entité. En collaborant avec d'autres institutions, la bibliothèque a la possibilité d'offrir des services étendus et d'étoffer son catalogue.

1.3 Les défis de la numérisation

Dans notre société du tout numérique, les systèmes informatiques et les standards d'aujourd'hui sont rapidement remplacés par de nouveaux standards. Cette évolution constante couplée à la masse informationnelle en perpétuelle croissance font naître une problématique complexe. L'un des grands défis de notre temps consiste à préserver nos données stockées afin que notre descendance garde une trace de nos activités en ayant accès à notre patrimoine.

L'obsolescence des formats de fichiers force les institutions à procéder à des migrations de leurs données de manière périodique. De plus, les supports de stockage physiques tels que les disques durs, les DVD et autres ont une durée de vie limitée. Parallèlement à cela, il n'est pas anodin d'éprouver des difficultés pour la recherche de fichiers informatiques mal référencés. Ici les métadonnées jouent un rôle très important. Malgré le coût important que nécessite la mise en place de ces « données sur les données », cette étape se révèle être primordiale pour l'établissement d'une conservation pérenne de l'information.

Les capacités insuffisantes de stockage, la perte de données, les données devenues illisibles ainsi qu'une expertise insuffisante sont les principaux problèmes rencontrés

par les institutions. Pour palier ces contraintes, il est nécessaire d'élaborer une politique de conservation numérique correspondant à son statut. L'importance d'élaborer une politique claire de nommage des fichiers, de formats, de stockage et de migration permettra de minimiser les contraintes liées à cette problématique.

La collaboration avec le service informatique est donc primordiale. Les systèmes informatiques doivent être opérationnels aussi bien pour les professionnels de l'information que pour les usagers. Les informaticiens ont pour tâches de gérer les fichiers sous différents formats, maintenir les performances réseaux tout en surveillant l'état des supports de stockage.

Le Référentiel général d'interopérabilité (RGI) propose des recommandations pour favoriser l'interopérabilité. Ce référentiel décrit un ensemble de normes et de bonnes pratiques utiles aux institutions publiques françaises. Ce dernier s'appuie sur la normalisation internationale (ISO, UIT) ainsi que sur des recommandations d'autres organismes (W3C, IETF, OASIS). La politique d'archivage, les schémas d'échange de données pour l'archivage ainsi que des préconisations en matière de formats sont traités dans ce référentiel.

1.3.1 Les formats

La bonne sélection de format joue un rôle prépondérant au niveau de la qualité et de la préservation à long terme des documents. Le choix des formats est donc une étape essentielle dans le processus de numérisation. Pour le codage et décodage des formats, des logiciels spécialisés entrent en scène. Ces logiciels (Adobe Photoshop, ASSY QuickScan Pro, etc.) peuvent être libres ou propriétaires et certains formats secrets comme Microsoft Word. Si ces logiciels devaient disparaître dans un futur proche, les données stockées peuvent devenir inaccessible. Dans une situation où le processus de codage est secret et que seul le logiciel d'origine est capable de déchiffrer les données, les pertes peuvent être dramatique pour une institution n'ayant pas anticipé cette problématique.

Le format doit être sélectionné selon différents critères. Il doit être ouvert, c'est-à-dire largement documenté et accessible. Il doit être utilisé par un grand nombre d'utilisateurs afin d'éviter une obsolescence prématurée. Il doit être normalisé et indépendant des autres formats et plateformes[3].

[3] Tableau récapitulatif des principaux formats de fichiers : Voir annexe 2

Pour information, la Bibliothèque du Congrès et la Bibliothèque royale des Pays-Bas utilisent le format JPEG 2000. Ce format est un bon compromis pour la conservation, cependant certains visualiseurs ne supporte pas ce format et peut donc être un frein à la diffusion.

Il est à noter que la Bibliothèque du Congrès propose une étude consacrée à l'évaluation des formats sur son site internet.

1.3.2 Les supports de stockage

Il existe différents supports pour le stockage des données, les supports optiques (CD, DVD) et les supports magnétiques (bandes, disques). Malheureusement, il n'existe pas encore de solution pérenne pour cette problématique. En effet, les supports sont fragiles et se détériorent au fil du temps.

Afin de limiter au mieux ces contraintes, il est important de veiller à la robustesse des supports. Le produit sélectionné doit être diffusé par un nombre important de fournisseurs et si possible être normalisé afin de minimiser les risques d'obsolescence. Par ailleurs, les conditions de stockage influent grandement sur l'état du support. Un environnement stable et contrôlé prolongera sa durée de vie. En règle générale, l'espérance de vie d'un support de stockage n'excède pas 10 ans.

1.4 Les aspects techniques de la numérisation

> *« La numérisation fait référence au procédé de traduction en bits d'un document d'information, comme un livre, un enregistrement sonore, une photo ou un vidéo. Les bits sont les unités fondamentales d'information d'un système informatique. Transformer une information en ces chiffres binaires est appelée numérisation. Ce procédé de numérisation peut s'effectuer grâce à diverses technologies existantes. »*
>
> (Source : http://bibliodoc.francophonie.org/article.php3?id_article=197)

En d'autres termes, la numérisation est la conversion d'une information présente sur un support physique en codage numérique sur un ou plusieurs bits, lisible par un programme informatique. Ce procédé aboutit à une image ou à une information textuelle. La conversion en mode texte du mode image peut se faire automatiquement par un logiciel de reconnaissance optique des caractères (OCR) ou par saisie manuelle.

La numérisation implique la création de bases de données multimédia contenant des références bibliographiques et différents types de données (images, son, vidéo).

Combinée au web et à l'internet, cette technologie offre aux usagers un large accès au patrimoine scientifique et culturel mondial.

1.5 La terminologie spécifique à la numérisation

Le procédé de numérisation ne se limite pas à placer un document dans un scanner pour en récolter un produit prêt à l'emploi. Avant de décrire précisément les différentes étapes du processus, il est nécessaire de rappeler quelques bases terminologiques. Afin de décrire au mieux et de façon compréhensible pour les profanes les termes techniques inhérent à la numérisation, je me suis basé sur le didacticiel d'imagerie numérique de la bibliothèque de l'Université de Cornell (USA) et sur le dernier ouvrage de Thierry Claerr et Isabelle Westeel, « Numériser et mettre en ligne ».

Comme mentionné auparavant, cette terminologie est consultable dans le glossaire en annexe.

2. Guide de numérisation

2.1 Problématique d'un projet de numérisation

Avant de se lancer dans un projet de numérisation, il est nécessaire de savoir pourquoi et pour qui l'on numérise.

La numérisation et la mise en ligne des documents relèvent de plusieurs défis. Un défi technologique tout d'abord où il s'agira d'évaluer consciencieusement les prestations et choix techniques. Un défi économique à court, moyen ou long terme. Un défi juridique avec la problématique liée aux droits d'auteur. Un défi organisationnel lié à la gestion du front et du back-office. Un défi scientifique qui nécessite de connaître les collections.

Tout projet de numérisation demande une organisation minutieuse, des ressources humaines et financières à pourvoir ainsi qu'un planning à tenir.

2.2 Les étapes du processus de numérisation

Tout projet de numérisation suit une démarche en plusieurs étapes.

2.2.1 Les objectifs

Il s'agit tout d'abord d'identifier les objectifs visés par la numérisation. Il est primordial de savoir pour quelles raisons un projet de numérisation doit être lancé. Les objectifs à

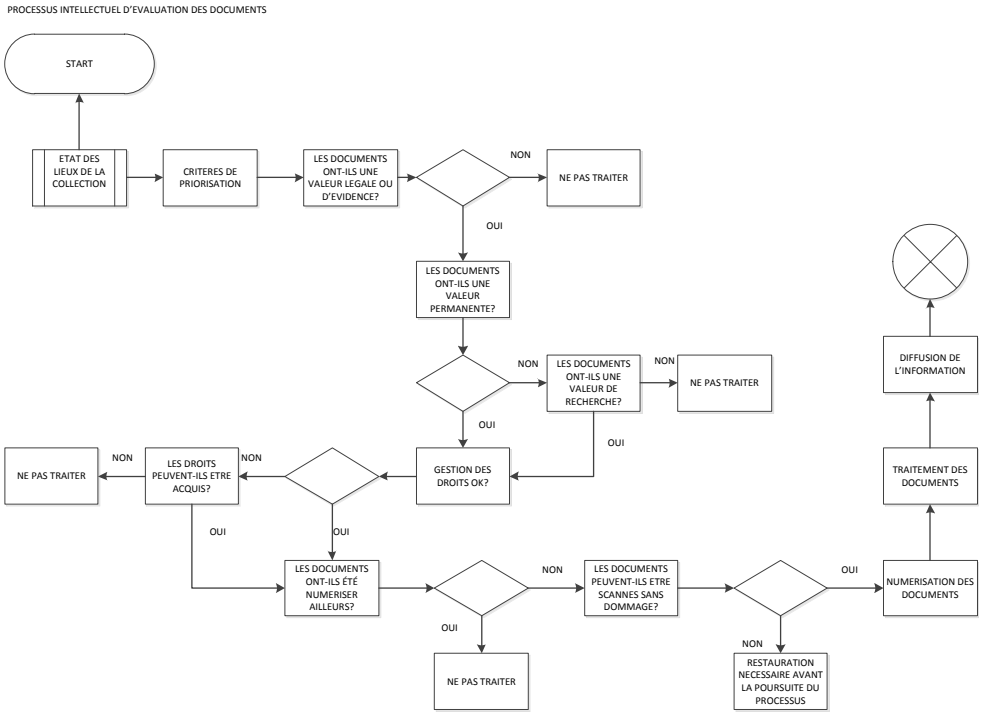
atteindre doivent être dûment documentés et les améliorations post numérisation clairement définies. Comme nous l'avons vu précédemment, cette technologie offre de nouvelles pratiques d'utilisation des documents permettant aux utilisateurs une consultation aisée et instantanée.

2.2.2 Les fonds documentaires

Avant de lancer un projet de numérisation, il est nécessaire de définir quels seront les documents à traiter. Dans notre cas, il ne semble pas judicieux de numériser en masse les collections des laboratoires car de nombreux documents sont tombés en désuétude au fil des ans. Effectivement, quel est l'intérêt de scanner un ouvrage sur Java 1.0. Il faut rappeler que les laboratoires n'ont pas de mission patrimoniale et n'ont donc aucun intérêt à conserver ce type de document. De plus, une sélection basée sur des critères intellectuels prenant en compte le contenu des documents permettra d'effectuer un tri important dans la masse documentaire devenue critique. Cette étape demande beaucoup d'énergie et est relativement longue. Par ailleurs, lors de la mise en place du processus de sélection, une attention particulière doit être portée sur les limitations légales. L'avantage est qu'à son terme, les documents à forte valeur ajoutée seront clairement répertoriés. Il est à noter que les critères de sélection sont souvent d'ordre thématique ou chronologique suivant les besoins des organismes.

Figure 1

Processus de priorisation des documents



2.2.3 Les attributs des documents

Dans un deuxième temps, il sera nécessaire d'étudier les documents d'un point de vue matériel. Ces derniers peuvent prendre différentes formes avec des caractéristiques bien distinctes. Il sera nécessaire de prendre en compte les dimensions de traitement des scanners et leurs facultés à traiter la couleur par exemple.

Le matériel documentaire se prête-t-il à la numérisation ? Le contenu informationnel peut-il être correctement saisi sous forme numérique ? Le format et l'état de conservation constituent-ils un obstacle ? Les intermédiaires tels que les microfilms ou diapositives sont-ils disponibles et en bon état ? Quelle est la taille et la complexité, en terme de diversité des documents trouvés, de cette collection ?

2.2.4 Coordination interinstitutionnelle

Un projet de numérisation demande un minimum de coordination. De plus en plus, les organisations visent à coordonner leurs efforts en matière de numérisation afin d'éviter les doublons et d'accélérer le processus. Afin de prendre connaissance des différents projets de numérisation en cours, la plateforme digicoord recense tous les travaux de

numérisation en Suisse. Elle permet également de mettre les institutions en contact pour des projets similaires. Pour les travaux déjà effectués, une consultation des principales bibliothèques numériques permet de prendre connaissance de l'offre documentaire.

2.2.5 Les infrastructures techniques

Les décisions concernant l'infrastructure technique nécessitent une planification particulière car la technologie de l'imagerie numérique évolue rapidement. Le meilleur moyen de minimiser les effets de l'obsolescence de ces outils est d'effectuer une estimation minutieuse, et d'éviter les solutions propriétaires uniques. Si les choix d'équipement correspondent parfaitement aux utilisations et résultats prévus, et s'ils sont synchronisés avec des calendriers réalistes, le retour sur investissement sera maximal.

Ce n'est qu'à cet instant que le travail de numérisation à proprement parler débute.

Avant de se lancer plus en avant dans ce travail, il est nécessaire de savoir si l'on désire conduire les opérations de notre propre chef ou de se tourner vers un prestataire externe.

Pour mener à bien un projet de numérisation interne, on veillera à sélectionner des produits standards et répandus sur le marché. Il est à noter qu'un bon soutien du fournisseur est fort appréciable dans de tels projets et ne doit donc pas être négligé. Il ne faut pas hésiter à investir dans du matériel de qualité afin d'éviter les problèmes d'obsolescence et d'incompatibilité. Les mises à jour du matériel informatique doivent pouvoir s'implanter aisément. Pour cela, il est important d'impliquer le personnel technique dès le commencement du projet. Une communication régulière entre les spécialistes de l'information et les informaticiens permet d'identifier plus rapidement les points faibles de la chaine de numérisation.

2.2.6 Les scanners

Ces appareils permettent de transformer un document en une image numérique. Le procédé de capture d'image est relativement simple. Un rayon lumineux balaie la surface d'un document puis, un capteur transforme cette lumière en un signal électrique qui sera interprété par l'ordinateur.

Nous pouvons diviser ces machines en trois catégories :

- Les numériseurs destinés à être mis en libre service dans une salle de lecture
- Les numériseurs de production dont les fichiers sont destinés à l'archivage
- Les numériseurs automatiques destinés à la production de masse

Le choix de l'outil dépendra fortement des besoins et attentes des institutions. Cependant, plusieurs paramètres sont incontournables pour sélectionner la machine adéquate.

Nous pouvons également les diviser en différents types :

- Les scanners à plat (le plus répandu car très facile d'utilisation)
- Les scanners à défilement (basés sur la même technologie que les précédents, ils en optimisent le débit)
- Les scanners à microfilm
- Les scanners à diapositives
- Les appareils photographiques numériques

Un échantillon de ces outils sera illustré dans les annexes correspondant aux prestataires de numérisation.

Tout d'abord il s'agira de choisir un numériseur adapté aux dimensions des documents à numériser. Il est nécessaire de prendre en considération la hauteur, la largeur et l'épaisseur de ces derniers. Les reliures ont également une incidence sur le choix de l'outil.

De plus, la résolution de l'appareil doit être conforme aux exigences de départ. Force est de constater qu'un balayage rapide des capteurs rend les couleurs moins exactes par exemple. Une attention particulière aux conditions d'éclairage ainsi qu'aux logiciels de traitement de l'image doit être prise en compte pour un rendu optimal.

Il n'est pas toujours aisé de réunir tous les facteurs en un même modèle. C'est pourquoi il sera inévitable de faire des compromis et de privilégier les critères les plus importants.

Nous pouvons estimer une productivité de plus de 1000 pages par jour en mode manuel pour des documents tels que des rapports, comptes-rendus de conférence et autres articles scientifiques. Une automatisation du processus peut tripler le résultat final dans des conditions optimales.

Voici un panorama non exhaustif des acteurs présents sur le marché actuel.

Tableau 1

Fournisseurs d'appareils de numérisation

marques	observations	Format
Assy-I2S	tourne page automatique avec ouverture à 180° du doc.	tout format
Kirtas	tourne page automatique avec ouverture partielle du document	Idem
Konica Minolta	Numériseur	A3.
Solar	Numériseur avec éclairage en option	A2.
Treventus	Tourne page automatique avec ouverture à 60° du document.	

2.2.7 Traitement de l'image/fichier

Le fichier nouvellement créé est ensuite soumis à différentes actions dans le but de corriger les imperfections, de le redimensionner, de le convertir dans d'autres formats, de le comprimer et d'y ajouter les métadonnées. Le traitement de l'image peut être intégré directement à la station de numérisation pour des opérations de redimensionnement par exemple. Pour des opérations nécessitant d'importantes ressources mémoires, un second ordinateur permettra de travailler plus efficacement. C'est le cas pour le traitement des images couleurs non compressées.

Par ailleurs, il sera nécessaire de repenser le nommage des fichiers. En effet, les répertoires et les schémas de nommage des fichiers sont rarement pertinents. Des précisions doivent être appliquées en particulier pour les collections importantes d'un point de vue quantitatif. Cette opération est délicate et lourde de conséquence pour l'exploitation ultérieure des fichiers. Pour ce faire, un plan de nommage doit être soigneusement mis en place en prenant en compte tous les cas de figure. Le code alphanumérique est une solution adaptée pour de nombreuses institutions. Il peut

contenir le numéro du fonds, la cote du document, le numéro de volume, le numéro de page, le type de fichier etc. Un tel code peut être très long mais permet d'identifier précisément le fichier et son contexte.

2.2.8 La conversion en mode texte

La conversion en mode texte permet la recherche plein texte sur les documents numérisés. Ce procédé permet donc de valoriser le contenu et de proposer de nouveaux services aux utilisateurs. Cette conversion peut se faire en mode manuel ou automatique par reconnaissance optique des caractères (ROC ou OCR en anglais). La technique d'OCR permet de situer et de reconnaître les chaînes de caractères dans une image et donc de faire la conversion des mots qui peuvent ensuite être utilisés pour faire une recherche plein texte. Cette conversion est assurée automatiquement par un logiciel et fait l'économie de la retranscription manuelle, beaucoup plus chère. Les mots et chaînes de caractères stockés dans un fichier texte peuvent être réutilisés pour une nouvelle mise en page, exploités dans une base de données, etc. Afin d'exploiter les résultats de l'OCR, la Bibliothèque nationale de France utilise un format basé sur XML et géré par un schéma, le format ALTO. Ce format a l'avantage d'être libre, ouvert et permet l'archivage et la réutilisation à long terme des données.

La qualité de ce procédé dépend fortement de la qualité du document numérisé. En effet, les images numériques doivent être suffisamment contrastées et redressées pour être traitées de manière optimale. De plus, les défauts d'impression, les polices de caractère très petites ou très grandes ainsi que les alphabets non latins sont des facteurs rendant la tâche plus compliquée.

Même en combinant plusieurs logiciels d'OCR sur un même texte, il est impossible, à l'heure actuelle, d'obtenir un taux de qualité de 100% (sans erreur). Pour cela, il est nécessaire de faire des ajustements manuels.

L'océrisation n'est pas une étape indispensable dans le cadre d'une évolution vers le numérique. Cependant elle permet d'offrir un certain nombre de fonctionnalités (notamment de recherche) aux utilisateurs. Ainsi, les fonctionnalités de recherche en texte intégral et de manipulation du texte (copier/coller, sélection d'une partie seulement du texte) ne pourront être proposées que grâce à une phase d'océrisation.

Cette étape se décompose en deux niveaux. Le premier est la réalisation d'un travail de reconnaissance des caractères à l'aide d'outils informatiques. Ces outils permettent d'obtenir un taux de reconnaissance variable en fonction des documents d'origine mais dans tous les cas insuffisants.

Le deuxième niveau concerne une relecture humaine des documents numérisés afin de vérifier l'ensemble du texte. Après relecture, on obtient un taux de fiabilité de l'ordre de 98%. Ce taux dépend également de la qualité de la numérisation effectuée. Bien entendu, le coût en temps de travail est loin d'être négligeable.

Cette étape peut devenir de plus en plus importante car les évolutions technologiques devraient permettre d'intégrer au travail d'océrisation, un travail de balisage. Sur la base de styles de présentation ou d'éléments de mise en page, on effectuera déjà une reconnaissance de certains éléments sémantiques (auteur, date, titre, etc.). La possibilité de mise en œuvre de ce type de fonctionnalités dépend du niveau de formatage des documents de base.

2.2.9 Le contrôle de la qualité

Le contrôle de qualité définit les différents critères de qualité requis pour les traitements des documents à numériser. Cette phase est primordiale lors du traitement de l'image numérique. Elle a été conçue afin de s'assurer que les attentes concernant la qualité sont remplies.

Cette phase peut être divisée en plusieurs étapes :

1. Identifier les produits

Il est nécessaire d'identifier les produits à évaluer dans un premier temps. Ceci concerne les images maîtres et dérivées, les impressions, les bases de données d'images et les métadonnées.

2. Développer une approche cohérente

Cette étape permet d'évaluer la qualité des produits dans l'optique d'établir un jugement. Il s'agit ici de définir si le produit est satisfaisant ou non. Des critères de qualité devront être choisis selon le type de document et la qualité d'image recherchée.

3. Déterminer un point de repère

Le point de repère permet de connaître sur quelle base on examine la qualité de l'image. En général on se base sur l'original.

4. Définition des objectifs et des méthodes de travail

Il est nécessaire d'identifier les objectifs de vérification. Est-il nécessaire de contrôler toutes les images ou seulement un échantillon ? L'évaluation sera-t-elle basée sur les originaux ou selon un regard subjectif ? La méthode de jugement doit être clairement documentée afin de distiller une évaluation de qualité.

5. Contrôle de l'environnement informatique

Les outils informatiques doivent être conditionnés afin de restituer une image fidèle du document original. La configuration informatique, le logiciel d'extraction d'image, les conditions de visionnage, la gestion de la couleur doivent être adaptés aux types de documents traités.

6. Evaluer la performance du système

Il s'agit de vérifier la performance du système avant la production. Cette phase concerne la numérisation interne et externe. Une vérification de la résolution, de la lumière, du bruit, de la reproduction couleur et autres sera menée.

7. Codifier les procédures d'inspection

Afin de faciliter les aménagements et les manipulations futures, les données de la qualité de contrôle doivent être codifiées. Les procédures, le personnel, les compétences requises, le matériel et les logiciels seront référencés pour faciliter le déroulement des opérations et servir pour la formation.

Il existe plusieurs méthodes de suivi de la qualité selon les besoins des institutions. Le contrôle exhaustif est très couteux et long. Le contrôle par échantillonnage qui se base sur des règles strictes. La mise à disposition d'outils de contrôle calibrés par le prestataire qui nécessite un investissement limité mais qui implique une maitrise limitée sur l'outil.

Par ailleurs, un contrôle qualité sous-traité peut être envisagé si les moyens le permettent. Les avantages principaux sont la fiabilité du contrôle (compétence du prestataire) et la possibilité de traiter des flux plus importants (échantillons à contrôler sur les livraisons sont de petite taille).

Il sera nécessaire, lors de la phase de test, de valider avec le prestataire les calibrages et ajustements des outils pour constituer un échantillon de documents représentatif de

la qualité souhaitée par le commanditaire. Il est à noter que cette phase de test conditionne fortement celle de la production. Un dialogue rapproché avec le prestataire doit permettre de préciser les procédures et les réglages techniques mentionnés dans le cahier des charges. Dès lors que cette phase de test est validée, la production peut débuter.

2.2.10 Les métadonnées

Leurs origines remontent au catalogage (indexation) de publications écrites. A l'ère du numérique, la naissance de nouvelles catégories de métadonnées a permis de gérer la navigation et la gestion des fichiers.

La description, l'indexation et la structuration des documents sont les principales tâches d'un spécialiste en information documentaire. Cette phase du projet de numérisation doit être mûrement réfléchie afin de maintenir les données et les métadonnées stables et pérennes. Les métadonnées (donnée qui décrit et définit une autre donnée) sont les garants de fonctionnalités de recherche et de navigation dans une bibliothèque numérique. Il est donc primordial de se baser sur des normes et des standards pour la création de ces métadonnées. La création et l'implantation de métadonnées sont des processus réclamant beaucoup de ressources. Il est nécessaire d'identifier les besoins de métadonnées en se basant sur les besoins des usagers et des responsables de la collection avant de procéder à la numérisation à proprement parler. Ceci afin de favoriser les processus de migration des données et d'éviter les problèmes liés à l'interopérabilité.

Il existe trois types de métadonnée qui caractérise un document numérique : les métadonnées descriptives, structurelles et administratives.

Les métadonnées descriptives sont établies dans un format normalisé (Dublin Core, EAD, MarcXML, etc.) qui permet les échanges. Elles regroupent les informations de contenu et d'identification d'un document (auteurs, titre, mots-clés, etc.).

Les métadonnées structurelles ou techniques donne des informations sur la version du document (date, format de fichier, taille, compression etc.). Elles permettent de connaître les relations physiques et logiques entre les fichiers tels que l'ordre d'affichage de ces derniers. Elles renseignent également sur l'environnement matériel et logiciel en vue d'une migration sur un autre système informatique.

Les métadonnées administratives informent sur les droits d'accès (droits d'auteur) et d'usage (impression, modification, etc.) ainsi que la préservation.

La création et la gestion des métadonnées s'effectuent manuellement (création d'un registre Dublin Core par exemple) ou automatiquement (création d'un index à mots-clés généré par la reconnaissance optique des caractères).

2.2.11 La mise en ligne des documents

La mise en ligne nécessite de mettre en place une architecture de recherche documentaire et de proposer différentes fonctionnalités aux usagers pour faciliter la consultation des documents. Les besoins des usagers se résument généralement en un accès rapide aux données dans une qualité d'affichage satisfaisante pour une consultation standard. Cependant, certains formats de fichiers ne sont pas optimisés pour une consultation via les navigateurs web. Cela nécessite de télécharger des plug-in ou autres applications qui, au final, s'avère être contraignant pour l'usager. Aujourd'hui, les organismes tentent de convertir à la volée des formats de fichier non-supportés en fichiers supportés par le navigateur afin de répondre aux demandes des utilisateurs.

La connexion au réseau doit être rapide, il est donc recommandé de porter une attention particulière à la taille des formats de fichiers ainsi que les connexions au réseau pour éviter de saturer les serveurs. Il est nécessaire également de soigner l'ergonomie de l'outil de recherche pour que l'usager puisse avoir accès directement au document sans devoir naviguer plus que de mesure pour trouver le lien.

En ce qui concerne le dépôt des documents sur la plateforme d'information de l'EPFL, nous y reviendrons dans le chapitre consacré à Infoscience.

2.2.12 La gestion des droits d'auteur

Rares sont les projets de numérisation pouvant faire abstraction des droits d'auteur et droits voisins. L'institution doit prendre connaissance des documents concernés par une telle protection et la durée de celle-ci si elle entend les diffuser à large échelle. Des autorisations d'usage doivent être préalablement négociées et il est primordial d'identifier les différents propriétaires des droits. Il faut savoir qu'une reproduction et une diffusion frauduleuse de données numériques peut mener à des poursuites judiciaires et de lourdes amendes.

Il existe des différences entre le droit suisse, le droit américain et européen. En suisse et en Europe, le délai est fixé à 70 ans après la mort de l'auteur pour qu'une œuvre tombe dans le domaine public. Aux Etats-Unis toutes les œuvres créées avant 1923 sont dans le domaine public. Et toutes les œuvres créées après le 1er janvier 1978 sont

protégées pendant une période correspondant à la durée de vie de l'auteur et 70 ans après sa mort.[4]

Il existe une autre problématique, celle des œuvres orphelines. Les œuvres orphelines sont des livres épuisés pour lesquels aucun détenteur de droit ne peut être identifié (estimation 40% des fonds de bibliothèques). Afin de traiter cette problématique, une solution européenne est mise en place, le projet ARROW (Accessible Registries of Rights Information and Orphan Works towards Europeana).

> *« L'objectif du Projet ARROW est de permettre à tout utilisateur, via une interface développée au niveau européen, de vérifier si une œuvre est disponible, épuisée ou orpheline, et d'obtenir des informations sur les détenteurs de droits. »*
>
> (Source : http://www.bnf.fr/documents/arrow_introduction.pdf)

En cas de doute et pour les questions liées aux droits d'auteur en Suisse, Pro Litteris et l'Institut fédéral de la propriété intellectuelle sont les organes compétents en la matière.

De manière globale, les prépublications d'œuvres scientifiques sont protégées par le copyright de l'auteur qui est donc libre de les diffuser à sa convenance. Cependant, dès lors que ses publications ont été validées par les pairs, l'auteur devra négocier avec l'éditeur les droits de cession de copyright qui reviennent à ce dernier.

2.2.13 La conservation des documents numériques

Tout projet de numérisation doit être accompagné d'une stratégie de préservation à long terme des données numérisées (importance des standards et formats non propriétaires). Avec la croissance exponentielle des données numériques, leur conservation est un véritable défi pour les institutions. Les formats de données concurrents sont nombreux sur le marché et plusieurs ont disparus au fil des ans. Il est donc important de développer des techniques adéquates de préservation du numérique, en multipliant les formats afin de faciliter une migration future par exemple. La conservation du document papier peut être une forme de sécurité également. Cependant, la gestion de conservation se voit alourdie. En effet, il n'est pas toujours envisageable de procéder de la sorte selon les organismes car certains d'entre eux numérisent leur documentation pour pouvoir ensuite se séparer des supports papier. Par ailleurs, le document numérique n'est pas insensible au vieillissement. Son code

[4] Loi fédérale sur le droit d'auteur et les droits voisins : http://www.admin.ch/ch/f/rs/2/231.1.fr.pdf

informatique (format) doit être interprété par un logiciel qui permettra la lecture via un écran de moniteur. Seuls les formats normalisés sont une garantie sur le long terme grâce à la description du mode de codage dans la norme. Le format Jpeg, très répandu dans le monde de l'informatique, n'est pas un format bénéficiant d'une norme officielle contrairement à ce que l'on pourrait croire. Il s'est imposé comme tel grâce à sa notoriété.

Afin d'anticiper au mieux les contraintes liées à la conservation, il est nécessaire de se poser les questions suivantes :

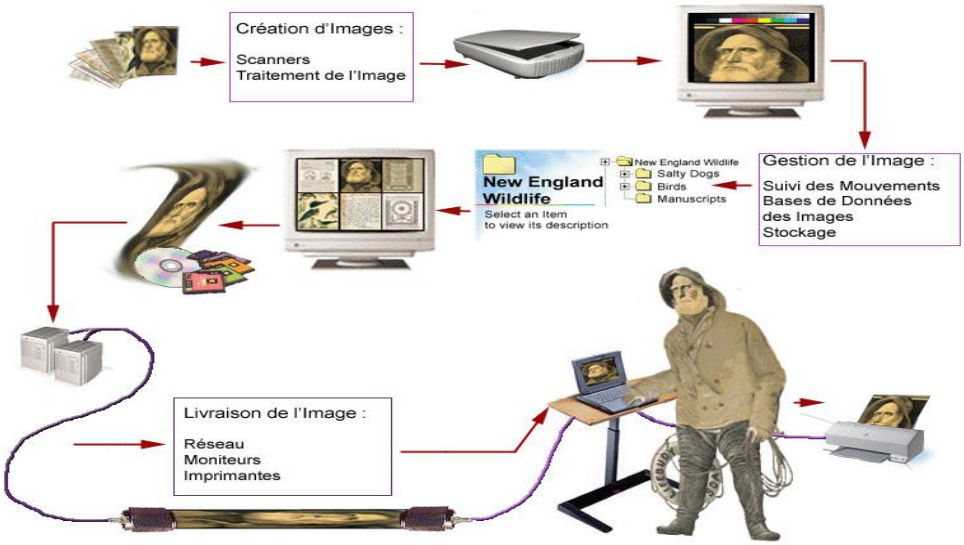
Le document court-t-il un risque pendant l'opération de numérisation ? L'image numérique peut-elle se substituer à la consultation des originaux, leur offrant ainsi une meilleure protection à la manipulation ? Est-il envisagé que la reproduction soit un moyen de remplacer les originaux ?

Tout projet de numérisation nécessite une planification minutieuse et ne laisse pas de place à l'improvisation. Il est à noter que les conséquences d'une gestion approximative peuvent être catastrophiques pour l'organisme. Repartir de zéro à la suite d'un enchainement d'erreurs humaines peut grever dangereusement le budget alloué à l'opération de numérisation.

Voici, pour conclure ce chapitre, un schéma de la chaîne de numérisation qui illustre parfaitement les différentes étapes du processus :

Figure 2

Chaîne de numérisation du didacticiel d'imagerie numérique de la bibliothèque de l'Université de Cornell



1. Capture numérique de l'image
2. Les types de scanner
3. Traitement de l'image/fichier
4. Gestion des fichiers
5. Bases de données d'images
6. Stockage
7. L'infrastructure réseau
8. Affichage sur moniteur

3. Les coûts de la numérisation

3.1 Méthodologie

Pour la rédaction de ce chapitre, je me suis basé sur le cours de Jean-Marc Rod, « numérisation patrimoniale », dispensé à la Haute école de gestion de Genève. Une étude menée par Benoît Eperon, ancien doctorant en sciences de l'information à l'Ecole nationale supérieure des sciences de l'information et des bibliothèques (ENSSIB), m'a permis de compléter les informations sur ce sujet. Je me suis également renseigné auprès d'institutions pour élaborer une estimation des coûts liés à la numérisation lors d'un processus interne, au sein de la bibliothèque.

3.2 Etat des lieux

Les coûts de la numérisation nécessitent une planification rigoureuse selon une méthodologie propre. Une bonne connaissance des objets à numériser est indispensable car les objectifs du projet vont déterminer son déroulement. Nous veillerons donc à mesurer l'atteinte des objectifs afin de pouvoir se prononcer sur l'amélioration apportée, notamment en terme d'accessibilité du document.

Comme nous l'avons vu précédemment, l'étape de numérisation consiste à transformer le document papier en un document électronique. Le coût de cette numérisation est fonction, d'une part, du matériel de base disponible et, d'autre part, du niveau qualitatif du résultat souhaité.

La réalisation de l'état des lieux des coûts de numérisation est inspirée de l'étude de Benoit Eperon, ancien doctorant de l'ENSSIB[5].

Au niveau des coûts, le prix matériel de la numérisation a fortement diminué. Le coût à la page a très nettement baissé ces cinq dernières années passant de plus de 1 CHF à 0.30 environ selon Silvio Corsini, conservateur de la réserve précieuse à la Bibliothèque cantonale et universitaire de Lausanne (BCU). De plus, le coût de fonctionnement de ce type d'équipement est faible car il ne nécessite pas de consommables tels que l'encre ou le toner. L'investissement principal réside dans les ressources humaines. Le travail de numérisation, et surtout celui de retouche des documents numérisés, est très coûteux en temps.

[5] Eléments pour l'appréciation des coûts de la numérisation : http://revues.enssib.fr/titre/2eco/2appreciation_couts/1numerisation.htm

Globalement, le coût de la numérisation réside principalement dans le volume horaire qu'elle nécessite. Comme nous l'avons vu, cette étape s'effectue généralement avec des exigences de qualité impliquant un certain niveau d'équipement et de compétence. En outre, les coûts liés aux droits d'auteur sont à prendre en compte dans un projet de numérisation. Suivant l'usage destiné aux documents électronique, il est primordial de s'entendre avec les ayants droit sous peine de poursuite judiciaire. Cette phase de négociation demande beaucoup de temps et implique des ressources financières importantes. Il est donc nécessaire d'identifier dès le départ les documents protégés.

La phase de sélection des documents permet de mettre en lumière les pièces nécessitant un travail de restauration avant d'être traité par le scanner. Pour les matériaux fragiles, le prix est important. En ce qui concerne le texte et les images, les coûts sont relativement modérés.

Pour un lot de documents homogène, les frais sont moindres. Ces derniers augmentent considérablement lorsqu'un lot contient différents formats aux exigences techniques spécifiques.

La planification d'un tel projet nécessite l'analyse des besoins et des attentes des différents protagonistes impliqués. Des entretiens doivent être menés pour préciser et cerner les points importants tels que la profondeur d'acquisition des documents et le format de mise en ligne.

Par la suite, il est nécessaire d'identifier les partenaires éventuels, les fournisseurs ainsi que les prestataires de service à impliquer dans le projet. Puis, une réflexion doit être menée au sujet de l'achat d'équipements spécialisés. Il est à noter qu'un tel investissement peut être pertinent si l'institution désire poursuivre le travail de numérisation une fois le projet terminé.

Suivant l'ampleur du projet, un chef de projet doit être désigné. Un projet d'envergure se situe à partir de 500'000 documents et s'étale sur plusieurs années. Cette personne doit être qualifiée et disposer de l'expérience nécessaire pour mener à bien le projet. En ce qui concerne le personnel, les ressources internes peuvent être suffisantes. Généralement, un projet de grande ampleur nécessite l'apport d'un personnel externe disposant des compétences requises.

Nous veillerons également à organiser correctement les locaux en y regroupant les ressources informatiques et les différentes machines. Les délais pour l'acquisition de

matériel sont parfois importants, dus principalement à la recherche et la sélection du produit répondant aux besoins de l'institution.

Les coûts liés à la gestion du projet durant son développement recouvre de nombreux domaines. La gestion du personnel doit être planifiée rigoureusement. Entretenir de bonnes relations avec les fournisseurs sera un atout appréciable tout au long du projet. La gestion des budgets, des contrats, des processus nécessite une attention particulière. Enfin, le contrôle qualité réclame un suivi particulier.

Par ailleurs, un contrat de qualité doit être dûment documenté pour des projets de numérisation pris en charge par un prestataire externe. L'absence d'un tel contrat peut avoir des conséquences graves sur la qualité finale ainsi que sur les délais de livraison. Ce document définit la nature du contrôle et qui en a la responsabilité.

Pour en revenir plus précisément au processus de numérisation, les coûts logistiques et relatifs aux étapes du processus de numérisation doivent être pris en considération.

Pour un organisme ne disposant pas de scanner automatique, où la numérisation de masse n'est pas concevable et possédant des ressources humaines à coût élevé, il faut compter une moyenne de 4 min/page (6h30 pour un ouvrage de 100p). Ceci prend en compte la numérisation, le traitement de l'image, la binarisation, la sauvegarde des fichiers et la création du PDF. Pour l'océrisation d'une page (sans vérification manuelle) une moyenne de 1min/p (donc 1h30 pour un ouvrage de 100p) est à prendre en considération. La facture moyenne d'un prestataire externe pour un travail de numérisation et d'océrisation est d'environ 0.344 euro/p sans compter le travail préalable (cahier des charges, récolement des fonds, contrôle qualité et feedback, finalisation des PDF, alimentation du site web, modification des notices du catalogue). Pour ce dernier, l'estimation monte à 0,486 euro/p. En comptant le travail préalable, la numérisation et l'océrisation on obtient un total de 0.83/p. Ces estimations sont tirées de l'expérience de numérisation menée en 2009 par la bibliothèque de l'ULB (Université libre de Bruxelles).

Ces estimations datent de 2 ans et les prix ont encore baissés depuis. Cependant, l'exemple de la bibliothèque de l'ULB démontre que les tarifs des prestataires de service ne sont pas négligeables et peuvent être un frein au projet de numérisation.

Nous pouvons en conclure qu'il est particulièrement pertinent de faire appel à un prestataire extérieur lorsque le lot à traiter représente un certain volume. Il est possible d'effectuer une économie d'échelle pour le transport des ouvrages et leur traitement et ceci est d'autant plus rentable que le lot est homogène et qu'il nécessite peu ou pas

d'intervention manuelle de l'opérateur. Le coût de numérisation proposé par les prestataires peut varier considérablement en fonction du volume, de l'état, de la fragilité et de la rareté du type de fonds à traiter, de la qualité attendue, des délivrables demandés.

3.3 Les coûts de la conversion

La conversion désigne la transformation de l'ensemble de la production scientifique des laboratoires dans des formats différents vers un format unique. Cette transformation permet par la suite une automatisation de certaines tâches (reconnaissance de styles pour le balisage, indexation, etc.).

Le coût de cette opération dépend donc évidemment des formats de départ et d'arrivée. Dans cette optique des consignes de mise en page doivent permettre un gain de temps. De plus, il faut s'assurer de disposer de personnes compétentes en interne ou en externe dans le maniement de différents formats. En effet, la conversion des documents en XML semble avoir été adoptée pour l'édition scientifique, ce qui nécessite des connaissances particulières.

Les coûts liés à la mise en place des systèmes de conversion automatisés et la conversion manuelle des formats non gérés par le système sont des éléments important à prendre en considération pour tout projet de numérisation. Les coûts initiaux peuvent être très élevés si l'on souhaite mettre en place une automatisation totale. En fonction des volumes à traiter, il peut s'avérer plus rentable de conserver une procédure manuelle.

Le préalable à toute opération de conversion est la définition du format dans lequel on désire stocker et exploiter les documents. Le choix doit se faire en fonction de la culture du public visé et du type de contenu des articles (certains formats sont plus destinés à la gestion des formules mathématiques par exemple). Il est également important de tenir compte des fonctionnalités que l'on souhaite mettre en place. En effet, des fonctions de liens hypertextes ou la finesse de la recherche possible (texte intégral, par mots-clés), la gamme de formats envisageables ne sera pas la même.

3.4 Les coûts de stockage

Le stockage des documents se fait à deux niveaux, le premier niveau est celui du document numérisé et le second celui du document converti.

Dans le premier cas, ce stockage se fait donc au format image, avant océrisation. Cette étape n'est pas anodine, notamment parce que les documents ou fichiers stockés constitueront le matériel accessible aux utilisateurs. Ce type de stockage doit être choisi en fonction de critères de sécurité et de conservation, mais aussi de capacité et de vitesse d'accès simultanés.
Dans le second cas, les documents stockés sont les documents électroniques ou les documents numérisés convertis. Cette étape de préservation vient comme une couche additionnelle au stockage des documents numérisés. Comme dans le premier cas, l'accès éventuel des utilisateurs à ces fichiers entraîne les mêmes critères de capacité et de vitesse d'accès.
Au niveau des prix, le stockage représente un coût principalement technique. Ce coût est relativement faible en raison de la baisse du coût de l'espace disque. A ce coût s'ajoute celui des serveurs et des connexions d'accès. Le niveau de sécurité et les critères de qualité de service influent fortement sur le budget global de cette phase (réplication des données sur plusieurs sites géographiques, installation dans des espaces sécurisés, et capacité du serveur).

Le stockage et la conservation des documents numérisés à ce niveau constituent un matériau de base dans le cas d'une conversion future dans un nouveau format. Si, à terme, le format de conversion choisi après océrisation se révèle obsolète, il peut être plus intéressant de recommencer la phase de conversion en partant du document numérisé.

3.5 Les coûts de la création des métadonnées

Les métadonnées peuvent être produites de deux façons distinctes. La première est la création manuelle par catalogage manuel des documents numériques ou papier. Cette indexation consiste à renseigner un certain nombre de champs dans une base de données permettant de décrire et d'indexer chaque document en fonction du niveau de granularité retenu. La deuxième est une création automatique à partir des balises insérées dans le texte ou reconnaissance automatique des éléments descriptifs (titre, auteur, source, date, etc.) dans le texte lui-même.

La première méthode de création manuelle peut s'appliquer aussi bien aux documents conservés au format image après la numérisation qu'aux documents électroniques en texte intégral. En effet, les métadonnées sont créées sur un autre support que celui du document initial.

La deuxième méthode nécessite une conversion des documents en XML afin de pouvoir les baliser en fonction de la DTD (Document Type Definition) préalablement définie. Les métadonnées intégrées à ce balisage peuvent être rassemblées dans une base de données. La création de la DTD représente une partie importante des coûts à l'heure actuelle en raison de l'absence de références dans ce domaine surtout si l'on souhaite développer une DTD spécifique et non pas utiliser celles déjà existantes. Il s'agit donc de formaliser avec précision l'ensemble des éléments d'un document comme par exemple un article ou d'un numéro de la revue (toujours en fonction de la granularité retenue) et leur organisation hiérarchique. Ces éléments seront ensuite repérés dans les documents au moyen de balises. Ce sont ces balises qui permettront l'extraction automatique des métadonnées. Ce processus de balisage est relativement automatisé. Il est possible de développer des systèmes de balisage automatique sur la base de styles ou d'éléments de mise en page. Cependant, un support humain de relecture, de vérification et de balisage (en l'absence d'éléments automatisables) est indispensable.

Les ressources techniques nécessaires sont relativement modestes car un poste informatique voir un spécialiste en information documentaire est suffisant pour effectuer le travail de création de la DTD et le balisage. Mais l'ensemble de ces tâches constitue un volume de temps de travail conséquent ainsi que des compétences relativement rares. C'est cette difficulté à disposer des ressources nécessaires en interne qui peut amener l'institution à opter pour une sous-traitance de cette étape. En effet, les coûts liés à la formation en interne sont importants. De plus, la productivité du collaborateur sera moindre dû principalement au manque d'expérience dans le domaine.

L'étape d'indexation consiste à classer et organiser l'ensemble des données recueillies au niveau précédent dans une base de données. C'est à partir de cette base de données que les articles seront recherchés par l'intermédiaire de l'interface proposée sur le site Internet.

3.6 Les coûts de la mise en ligne

Les différents aspects de la mise en ligne sont présentés ici dans le cadre de la création d'un site Internet. Les postes budgétaires présentés respectent une certaine chronologie dans la mise en place d'un site Internet.

3.6.1 Achat du matériel

L'achat de matériels et logiciels est un poste difficile à évaluer du fait de la grande diversité des situations possibles et de l'évolution permanente des produits. De plus le choix de l'environnement logiciel est lié au choix d'un certain nombre de standards de production ou de traitement. Cet environnement mouvant incite à ne pas négliger le poste « Maintenance du site ».

3.6.2 Hébergement et accès

L'hébergement du site Web peut se faire suivant deux options : un hébergement en interne dans l'hypothèse où l'entreprise l'université dispose du matériel et de la connexion Internet adaptés; sinon, l'hébergement se fait chez un prestataire extérieur. Le choix de l'une ou l'autre option doit se faire en tenant compte d'un certain nombre de critères : coût, facilité de mise à jour, présence des compétences en interne. Les orientations prises dans ce module conditionnent un certain nombre de modules notamment le recrutement de compétences ou le fonctionnement du site ainsi que les moyens d'accès réseau mis en place.

3.6.3 Mise en place et maintenance

Selon le cas, la création de l'infrastructure du site peut être externalisée ou réalisée en interne. De ce choix stratégique découle un grand nombre d'éléments. La différence en terme de coûts et de mode de fonctionnement lors de la phase de réalisation peut être importante. Quel que soit le mode de réalisation choisi, une formalisation des besoins doit être planifiée par un cahier des charges. Ce travail préalable représente un investissement en temps non négligeable.

La constante évolution logicielle implique d'anticiper des évolutions du site en terme d'interface ou de technologies employées. Cette fonction peut être couplée à une fonction de maintenance technique.

Fonctionnement du site

La mise en forme des publications en fonction des feuilles de styles ou/et l'intégration des éléments de structure conformes à la DTD nécessitent des ressources humaines conséquentes. Dans le fonctionnement du site il faut également inclure l'ensemble des éléments de gestion administrative et de ressources inhérents au site. Enfin, est inclue dans ce module une partie de l'activité de relecture.

4. Les prestataires de numérisation

4.1 Méthodologie

Afin de me faire une idée concrète des méthodes appliquées par les professionnels du domaine, plusieurs entreprises m'ont ouvert leurs portes pour une visite guidée. J'ai pu découvrir les machines, les méthodes de traitement des images, les produits finis en attente de livraison, etc. Je me suis équipé d'un appareil photographique pour la capture d'image des locaux. Ces clichés, accompagnés d'une brève description, peuvent être consultés dans les annexes de mon rapport. Je me suis limité à quatre organismes de Suisse romande dont un faisant partie de l'EPFL. Ce dernier ne pratique pas de travaux de numérisation à l'heure actuelle. Cependant, l'école polytechnique évolue de manière exponentielle ces dernières années et le service de reprographie pourrait se voir attribuer de nouvelles tâches dans l'avenir. J'ai donc décidé d'inclure ce service dans cette présentation.

Dans le cadre du cours de numérisation patrimoniale dispensé lors du 6e semestre de notre formation, mes camarades et moi-même avons eu l'opportunité de visiter les locaux de l'entreprise 4DigitalBooks durant une journée. Nous avons pu suivre les différentes étapes du cycle de numérisation. En outre, nous avons profité de la disponibilité des professionnels pour leur poser de nombreuses questions et de les photographier durant leurs tâches. Nous avons également testé par nous même les nombreuses machines et outils de numérisation.

La visite de l'entreprise Secur'Archiv à Genève n'a pas pu se dérouler comme souhaité. Je devais me rendre, accompagné d'une camarade de classe, dans leurs locaux de numérisation pour suivre une présentation globale de leurs prestations. Un empêchement de dernière minute m'a contraint d'annuler cette visite. Mon rapport ce base sur les informations récoltées par ma camarade et les échanges de courriels avec les professionnels.

4.2 Etat des lieux

A l'heure actuelle, il n'existe pas de service de numérisation au sein de l'EPFL, ni d'ailleurs de service d'archives doté d'une expertise ou d'une mission visant à la conservation numérique du patrimoine documentaire ou archivistique de l'EPFL. Chaque service de l'école procède de manière autonome suivant ses besoins. La bibliothèque de l'EPFL possède deux machines mises à disposition des usagers. La dernière acquisition est un scanner Zeutschel OS 12000, qui complète l'achat d'un

modèle plus rustique I2S E-scan, ce qui permet de disposer de 2 scanners de livre manuels. Par ailleurs, L'EPFL s'est équipée de scanners MFP noir et blanc et couleur. En 2009, des appareils multi-fonctions (MFP) ont été mis à disposition des étudiants et autres usagers à travers l'école. Ces emplacements spécifiques sont appelés PrintSpaces. Actuellement, le parc est constitué de dix WorkCentre 5665 noir/blanc et d'un WorkCentre 7665 couleur (situé au RLC).

Ces appareils, dotés d'un module de finition Microsoft Office, peuvent imprimer/photocopier en A4 ou A3, en recto ou recto-verso et agrafer des paquets allant jusqu'à 50 feuilles de papier.

En ce qui concerne le scan, il s'effectue par défaut au format PDF/A, à 300 dpi, en recto seul, destiné à l'OCR. Les documents sont donc indexés, c'est à dire que le texte est reconnu, ce qui permet d'effectuer des opérations propres au texte, comme le copier/coller ou la recherche.

Cependant, ces machines ne peuvent répondre aux besoins d'un projet de numérisation d'envergure. De plus, le centre de reprographie de l'EPFL ne prend pas en charge les travaux de numérisation, sauf cas exceptionnels comme nous le verrons plus bas.

Au vu de ce constat, un projet de numérisation d'envergure devra probablement être finalisé avec un partenaire externe. Il est donc intéressant d'établir un panorama des principaux prestataires de service en Suisse romande. Pour cela, j'ai rencontré des professionnels de différents organismes de numérisation en Suisse occidentale.

4.3 Présentation des organismes

4.3.1 Atelier de numérisation de la Ville de Lausanne (visité le 12 avril 2011)

4.3.1.1 Présentation de l'atelier

Il y a neuf ans que l'atelier de numérisation de la ville de Lausanne a été fondé. L'atelier a déménagé récemment au Service d'Organisation et Informatique (SOI) de la ville. Il était basé auparavant au Musée historique (Ancien-Evêché).

Sa mission première est de photographier et de numériser divers types de documents iconographiques. Les usagers sont principalement des historiens, des étudiants et des journalistes.

L'atelier traite tous les supports iconographiques des musées communaux et du Fonds des arts plastiques. Les professionnels effectuent des prises de vue et s'occupent de la gestion des couleurs. Tous les collaborateurs ont une formation en photographie. La clientèle se compose quasiment exclusivement d'institutions communales sauf rares exceptions. En effet, l'atelier n'a pas pour habitude de travailler pour des entreprises privées. Cependant, il peut mettre en location sa machine ID2S.

L'atelier possède un numériseur grand format (ID2S), deux scanners Heidelberg et un traceur graphique. De plus, une salle de photographie, munie d'une caméra Sinar (pour les prises de vue d'objets en 3D tel que sculptures, œuvres d'art, argenterie etc.), de nombreuses lampes et réflecteurs, viennent compléter le parc technologique. Tout ceci est piloté par ordinateur.

4.3.1.2 Les outils technologiques

Voici, en détail, les machines présentes dans les locaux de l'atelier. Je n'ai pas pu les voir en action, cependant Olivier Laffely, responsable de l'atelier, m'a présenté chacune d'elles.

- ID2S : dédié aux très grands formats. La caméra de plus de 14 millions de pixels permet un scan minutieux des documents. L'éclairage se fait par LED afin d'agresser le moins possible les œuvres fragiles. La table est amovible, ce qui est utile pour les grands formats difficiles à manœuvrer. De plus, le numériseur est muni d'une table d'aspiration pour les documents non-planes. Un ordinateur relié directement à la machine permet de traiter les images grâce au logiciel Photoshop. La gestion des couleurs est paramétrable via cet outil.

- Nexscan (Heidelberg) : machine utilisée en imprimerie. L'atelier en possède 2 qui ont plus de 8 ans. L'entreprise allemande ne fabrique plus de numériseur et ces 2 machines seront bientôt remplacées. Elles sont utilisées pour les petits formats et ceux de taille standard (diapositives, photographies, livres, périodiques etc.)

- Traceur graphique : permet l'impression de plans, d'affiches, panneaux d'exposition etc.

- Caméra Sinar: situé dans le studio de photographie, elle propose une résolution de plus de 30 millions de pixels. Il permet le calcul des bascules, une visée télémétrique etc. Cet appareil est utilisé pour les clichés d'objets en trois dimensions tels que l'argenterie, les sculptures et les œuvres d'art.

4.3.1.3 Le traitement des fichiers

Le format des fichiers ainsi créés est le TIFF, très lourd (400M le fichier) mais qui restitue très bien les couleurs et est très souple à l'utilisation. Les fichiers sont ensuite gravés dans une « librairie optique » (base de données d'images). La résolution de l'image est poussée au maximum pour respecter le mieux possible l'intégrité du document numérisé. Cependant, une fidélité chromatique à 100% n'est pas possible. Les métadonnées sont intégrées manuellement via Photoshop. La vérification de la qualité se fait avec ce logiciel. Une attention particulière est donnée au contraste, cadrage, couleurs et à la compression. Les fichiers comprimés sont au format LZW (sans perte). Les professionnels vérifient également si la cote de l'objet correspond bien au document et si la correspondance avec la base de données est fonctionnelle. Ils vérifient ensuite la pertinence des métadonnées. Ce travail de vérification est lourd et chronophage mais indispensable.

4.3.1.4 La base de données

Une base de données regroupe tous les documents scannés. Un masque de recherche divisé en plusieurs champs (institution, titre, auteur etc.) permet de cibler précisément un document. Afin d'utiliser les images aisément, les musées peuvent appauvrir celles-ci d'avantage. A la base, un fichier TIFF n'est pas adapté à la consultation en ligne car trop lourd en terme de volume.

4.3.1.5 Conclusion

En guise de conclusion, il est à noter que l'atelier de numérisation traite quasiment exclusivement avec les partenaires de la Ville de Lausanne et les institutions communales. Une collaboration avec l'EPFL ne semble donc pas d'actualité. Cependant, Olivier Laffely ne ferme pas les portes pour autant mais la décision finale appartient à la Ville, souligne-t-il.

4.3.2 4DigitalBooks, ASSY SA (visité le 13 avril 2011)

4.3.2.1 Présentation

Ivo Iossiger est ingénieur diplômé de l'EPFL et président de 4 DigitalBooks à Ecublens. Conscient des problèmes d'archivage des données dans certains instituts et des moyens archaïques mis en place, Ivo Iossiger et ses collaborateurs ont eu l'idée de créer un concept nouveau profitable à tous. Lorsque la BNS a acquis un scanner pour la numérisation des documents, son fonctionnement était de type manuel. Il fallait donc créer une machine automatisée. A l'époque, il n'existait rien de comparable sur le

marché. C'est à Saint-Aubin (NE) qu'Ivo Iossiger débute ses travaux. Trois années de dur labeur, et l'équipe est finalement parvenue à l'association parfaite de l'informatique et de la mécanique.

En 2001, Ivo Iossiger reçoit le Prix de la meilleure invention, décerné par le Wall Street Journal.

C'est en 2004 que débute la production de service de numérisation pour des clients. L'entreprise décroche son premier contrat avec l'Université de Stanford (USA). Aujourd'hui, 4digitalbooks traite 3 millions de pages par année.

4.3.2.2 Le cycle du document

Les documents entrants sont stockés dans une pièce prévue à cet effet. Des étagères et des coffres forts accueillent les documents selon leur état et leur valeur. Différents supports sont traités (peintures, diapositives, journaux, livres etc.) A leur arrivée, un fichier Excel regroupe des données sur l'état, le format et la composition du document. Un numéro d'identification est attribué à chaque document ainsi qu'une grille de travail avec les recommandations du client, l'arborescence, le nommage des fichiers, les liens avec les bases de données etc. Une fois numérisé, un intercalaire est glissé dans le document.

4.3.2.3 Les outils technologiques

Tout au long de notre visite, nous avons eu l'occasion de tester différentes machines. Tout d'abord nous nous sommes intéressés à un scanner de type photocopieuse qui fonctionne par balayage. On place les feuilles à numériser sur le dos de l'appareil puis il les avale à une cadence soutenue. Cet outil est utilisé principalement pour les documents d'archive. Grâce au logiciel QuickScan pro, une multitude de paramètres peuvent être saisis (format, recto-verso, niveau de gris, résolution, type de fichier etc.). La résolution maximum est de 300 dpi et la machine gère les niveaux de gris et la couleur. Avant de traiter un document, on crée un nouveau lot. Cela permet, par la suite, de gagner du temps lorsque des documents similaires devront être traités. Finalement, une vérification manuelle sommaire des fichiers bruts (RAW) est effectuée avant de procéder au traitement des fichiers. Une fois numérisé, les pages ne sont pas toujours bien droite, on devine parfois les caractères au verso. Le logiciel maison Page Improva permet de réduire voire éliminer tous ces défauts. Un travail de nettoyage est effectué par zone du document. Recadrage, modification du gamma, pli de reliure sont traités pour remettre le document au goût du jour tout en préservant son intégrité.

La deuxième machine que nous avons pu tester est destinée aux grands formats tels que des livres, des cartes, des peintures etc.). Le scan s'effectue de manière linéaire. Un bras articulé déplace la caméra de gauche à droite. Il est possible de changer la mise au point, l'ouverture du diaphragme, la gestion des couleurs etc. L'éclairage se fait encore par néon bien que la tendance soit au LED, plus stable. Pour le traitement des plans, l'entreprise collabore avec l'atelier de numérisation de la Ville de Lausanne qui possède une machine adaptée.

La dernière machine testée est de type manuel munie de 2 caméras aux optiques Zeiss 50 mm. Ici le format maximum est A4 et l'opérateur doit tourner les pages du livre et lancé le scan à chaque page. Cette machine est destinée aux ouvrages délicats qui ne peuvent pas être traités de manière automatique.

Pour clore notre visite, une présentation d'une machine appartenant à la Digitizing Line, la DL 3000, nous a été proposée. Véritable bijou de technologie, cet outil ne nécessite pratiquement aucune intervention humaine. Elle traite 3000 pages par heure. La machine détecte automatiquement le format du document. La résolution peut monter jusqu'à 800 dpi.

4.3.2.4 Conclusion

Cette visite m'a permis de prendre conscience concrètement de la complexité du travail de numérisation au sein d'une entreprise de production. L'automatisation des processus et la production de masse tranche totalement avec l'aspect plus « artisanal » de l'atelier de numérisation de la Ville de Lausanne.

De part sa proximité géographique et la qualité de ses services, 4DigitalBooks se positionne comme un partenaire de premier choix pour l'EPFL.

4.3.3 SecurArchiv SA

4.3.3.1 Présentation

Fondée à Genève en 1985 par Jean-Jacques Borgstedt, PDG du groupe Pelichet, Secur'Archiv SA est une entreprise active dans la conservation et la gestion d'archives et la protection de données professionnelles. Elle dispose aujourd'hui de dix centres d'archivage et emploie plus de 30 collaborateurs.

L'entreprise est localisée sur plusieurs sites: Genève, Lausanne, Bâle, Zürich et Berne. Cependant le centre de numérisation se trouve à Genève. Secur'Archiv est active dans la conservation et la gestion dynamique d'archives papier, notamment l'entreposage, la surveillance et la traçabilité des documents. Elle offre également un service de

livraison adapté aux besoins des clients. L'entreprise met à disposition de ses clients différents logiciels pour la gestion d'inventaire, le traitement des commandes, la numérisation et le scanning des documents. Il est à noter que le service de numérisation peut se dérouler chez le client ou dans les locaux de Secur'Archiv. L'entreprise dispose d'un centre de numérisation à Genève sur 150 m^2 regroupant une gamme complète de scanners haute performance (scanners à livre, à plan, etc.).

Secur'Archiv archive des documents papier et électroniques sur disque dur pour ses clients. Pour ceci, ils possèdent de grands entrepôts sur leurs différents sites afin de stocker les kilomètres de documents. Ils utilisent plusieurs méthodes de compactage, variant selon le degré d'utilisation et le type de document. Les documents peuvent être entreposés dans des containers maritimes, dans des boites en carton ou dans des compactus.

La clientèle est hétérogène. Elle est composée d'entreprises privées, d'organisations internationales, ou d'institutions publiques suisses comme la Ville de Genève.

4.3.3.2 Clauses contractuelles

Pour passer contrat avec SecurArchiv, il est nécessaire de créer un cahier des charges qui détermine les rôles de chacun, les spécifications concernant la numérisation, les délais, etc. Ce cahier des charges sera le document de référence déterminant les rapports entre partie. Celui-ci doit être rigoureux et complet. Le coût final de numérisation variera en fonction des éléments inclus dans le cahier des charges, de la quantité de document et des délais souhaités par le client. Plus les délais sont cours, plus le prix est élevé.

4.3.3.3 Conclusion

A l'instar de 4DigitalBooks, l'entreprise traite de grandes quantités de document à une cadence industrielle. De plus, l'entreprise prend en charge les images sous forme de diapositive, ce qui n'est pas le cas de son concurrent. En ce qui concerne les prestations globales, Secur'Archiv prend en charge le transport et le traitement des lots et les professionnels se déplacent chez le client si les documents ne peuvent être transportés. Au vu de la qualité de ses services, Secur'Archiv est un partenaire à prendre en considération pour l'EPFL.

4.3.4 Reprographie EPFL (23 MAI 2011)

4.3.4.1 Présentation

Le service de reprographie fait figure d'exception dans cette présentation car elle ne propose pas de prestation de numérisation. Cependant, elle propose de nombreux services pour la communauté EPFL et pourrait bien un jour se voir attribuer des tâches de numérisation.

Créé en 1975, le service de reprographie a pour mission de soutenir l'enseignement et de valoriser la recherche à l'intérieur comme à l'extérieur des murs de l'école en assumant les tâches liées à l'impression de divers types de documents produits par les enseignants et l'institution. Cette entité emploie neuf collaborateurs issus des métiers de l'impression (typographes, relieurs, etc.).

Son activité se concentre essentiellement sur la reproduction sur support physique et ce service ne possède pas de ressources de numérisation à proprement parler. En effet, selon Thomas Reynaud, son responsable, le service de reprographie ne possède ni le personnel, ni les ressources financières pour développer un tel service et il n'y a pas actuellement de projet allant dans ce sens à moyen terme. Le service possède cependant un vieux scanner (format A3 maximum et prise en charge de la couleur) pour des tâches de numérisation ponctuelles. Il n'est donc pas envisageable pour le service de reprographie de pouvoir se lancer dans un projet d'envergure en l'état actuel.

4.3.4.2 Les outils technologiques

Les ressources matérielles se composent notamment de deux lignes Xerox pour l'impression des thèses, un offset pour le traitement des couvertures et d'un plotter pour les grands formats. La reprographie a acquis récemment deux copieurs couleur de dernière génération (Konica-Minolta C6500) et d'une nouvelle machine Dynic pour la réalisation de couverture enveloppante.

4.3.4.3 Conclusion

Le service de reprographie ne se profil pas en tant que prestataire de numérisation à l'heure actuelle. Cependant, la dimension des locaux et les compétences techniques des collaborateurs offrent de nombreuses opportunités pour ajouter la thématique de la numérisation au cahier des charges.

4.4 Tarification

4.4.1 Méthodologie

Afin de sonder le marché et de comparer les offres tarifaires des différents prestataires, j'ai élaboré une sélection de documents plus ou moins représentatifs des fonds à numériser des laboratoires. Ce document a pour objectif de tester la rapidité de la prise en charge des fonds, les délais de livraison et le montant de la prestation globale de numérisation. Lors de ma demande, il a été précisé qu'un devis n'était pas souhaité mais une estimation des coûts. Le fait de ne pas avoir de données exhaustives sur l'ensemble des collections ne me permettait pas d'établir un « appel d'offre » en bonne et due forme.

A l'heure de mettre sous presse, deux prestataires ont accepté de répondre à ma demande. Il s'agit de l'entreprise Arcplace AG sise à Genève et de 4DigitalBooks à Ecublens. Cependant, le service client d'Arcplace AG ne m'a pas communiqué leurs offres malgré un entretien téléphonique et plusieurs contacts par courriel.

4DigitalBooks a répondu, par l'entremise de M. Rod, responsable du « Business Developement Service » pour l'Europe, à mon appel d'offre sommaire. Le compte rendu complet est disponible dans les annexes. Par ailleurs, le calcul des coûts, pour la prise en charge de la documentation des laboratoires étudiés, se base sur les données transmises par M. Rod.

En ce qui concerne Secur'archiv S.A., ils ne désirent pas se prononcer sans avoir pu visiter les collections. Cependant, le responsable qualité de l'entreprise m'a fournit quelques chiffres afin de sonder les prix du marché.

M. Lang, responsable assurance qualité, indique que les tarifs de numérisation s'étalent de 10 centimes à 2 CHF par page (voire plus) selon le travail à effectuer. Au prix de 10 centimes, il s'agira de grandes quantités de pages A4 en noir et blanc, sans agrafes, et sans indexation, à numériser sur des scanners de production. En ce qui concerne la numérisation manuelle sur des scanners à livres, les prix s'étalent entre 0.70 et 1.50 CHF par page. Pour les grands formats (plans, affiches, etc.) une tarification forfaitaire est proposée en fonction du nombre de plans ou d'affiches. Par exemple, 450 CHF pour 25 pièces.

De nombreux facteurs viennent encore alourdir la facture finale tels que le volume du fonds, le traitement recto verso, la préparation (agrafes, trombones, type de papier, etc.), la gestion de la couleur, etc.

L'indexation peut aussi avoir un impact économique non négligeable, en fonction du nombre et de la complexité des informations à renseigner, et des possibilités ou non d'indexation automatique.

5. Présentation des laboratoires et unités étudiés

5.1 Etat des lieux

Ce chapitre présente les quatre laboratoires sélectionnés et leur gestion documentaire, le Service académique et la collection de tirés-à-part de mathématique étudiés au cours de mon travail. Chaque entité à ses spécificités et son mode de fonctionnement propre à sa situation documentaire. Le service académique apparaît comme une exception dans mon rapport. Pour faire face à la masse documentaire exponentielle des documents d'inscription et de gestion des étudiants, une solution de gestion électronique des documents a été mise en place. Ce service est présenté ici comme un exemple pour illustrer la démarche de numérisation.

1. Tirés-à-part de mathématique (Rolex Learning Center)
2. Documentation de la Communauté d'étude sur l'aménagement du territoire (CEAT)
3. Documentation de l'Ibeton (Enac)
4. Documentation de l'Icom (Enac)
5. Documentation du laboratoire des machines hydrauliques (LMH)
6. Gestion des archives courantes au Service académique (SAC)

5.1.1 La collection des tirés à part de mathématique

Propriétaire de la collection: Bibliothèque de l'EPFL

Localisation : Salle des compactus au sous-sol du RLC, rangée A 117 – A 118

Nombre de documents: 6000 à 8000 documents.

Utilisateurs/usagers: professeurs.

Droits: éditeurs commerciaux / auteurs.

Présentation de la collection

Un « tiré à part » est une impression séparée d'une partie d'ouvrage ou d'un article de périodique, relié à part avec une pagination propre en un petit livret. Cet article représente le résultat d'un travail de recherche spécifique qui constitue par conséquent, une unité documentaire à part entière.

Les documents de la collection de l'EPFL datent des années 70 et antérieures. Ce sont principalement des thèses et des tirés à part provenant d'auteurs extérieurs à l'EPFL. Les documents se présentent sous forme de fascicules et de feuillets. De prime abord, les tirés à part ne représentent pas d'intérêt particulier. Ce sont des articles scientifiques (reprints) tirés de revues spécialisées. Après une analyse sommaire du fonds, la plupart des documents sont disponibles dans d'autres institutions selon David Aymonin.

40% de la collection est catalogué dans une base de données 4D[6] (avant la centralisation à la bibliothèque de l'UNIL). La licence a été arrêtée et l'import des données dans Aleph[7] n'a pas pu être effectué (problème d'interopérabilité avec le format MARC). Les priorités visant d'autres collections, les tirés à part ont été mis de côté, la collection étant morte depuis les années 70.

Les tirés à part sont conservés dans des boîtes d'archives noires dans les compactus au sous-sol de la bibliothèque. Classés par ordre alphabétique d'auteur, un inventaire sommaire (dans chaque boîte) décrit le contenu des divers documents.

Le fonds ne contient pas de documents manuscrits ni de dépliants. Les documents semblent pouvoir être numérisés aisément sans risque de dégâts, cependant après vérifications, il s'avère que les tirés à parts les plus anciens souffrent d'un jaunissement avancé du papier (papier acide) et un décollement de la reliure. Dans certains documents, des fiches bibliographiques ont été insérées. Cela prouve la présence d'un catalogue papier.

Les notices bibliographiques de la collection n'étant pas mises en ligne, quelques professeurs ayant connaissance de ce fonds viennent consulter les documents. Cependant, ces visites sont très rares.

Types de documents : articles de périodique et thèses.

Formats des documents : A4 et A5.

Nombres de documents : entre 6000 et 8000 pièces.

Etat de la collection : se dégrade dangereusement.

Consultation : rare.

---

[6] 4D est un environnement de développement intégré. http://www.4d.com/fr/

[7] Aleph est un Système intégré de gestion de bibliothèque développé par la société Ex Libris. http://www.exlibrisgroup.com/fr/category/ILSOverview

Tableau récapitulatif des documents présents dans le fonds des mathématiques

Type de documents	Quantité
articles de périodique (reprints)	?
Thèses	?

5.1.1.1 Contraintes techniques

L'absence de catalogue pour cette collection rend la tâche d'analyse de celle-ci difficile à appréhender. Afin de remettre la main sur le catalogue, une solution serait de réactiver la licence et vérifier si les notices sont encore valables aujourd'hui sur le nouveau système. Conjointement avec Mme Guéritault, bibliothécaire en charge de cette collection de tirés à parts, nous avons tenté d'accéder au catalogue sans résultat. Nous avons donc fait appel à un informaticien, qui avait pour charge à l'époque la gestion de la licence 4D. Malgré plusieurs essais, notamment la lecture du disque de sauvegarde, les fichiers ne peuvent être consultés. Le numéro de licence n'étant plus en possession de la bibliothèque, la seule solution est de contacter l'entreprise 4D pour débloquer la situation. Cependant, le renouvellement de la licence demande un investissement onéreux pour un résultat aléatoire. En effet, les notices bibliographiques seront-elles compatibles avec le système actuel ? Lors du déménagement des bibliothèques au sein du Rolex Learning Center, Mme Guéritault avait rencontré des problèmes de migration des données dans Aleph (système intégré de gestion de bibliothèque). Il était impossible d'exporter les notices de Bibliomaker dans ce système.

Dans ces conditions, mon projet se base sur une estimation de 6000 documents par défaut.

5.1.1.2 Recommandations pour la numérisation

Idéalement, il serait nécessaire de réactiver la licence 4D pour consulter le travail de catalogage déjà réalisé sur cette collection. Force est de constater que l'absence de catalogue rend la tâche ardue pour un travail de numérisation. Avant de se lancer dans un tel projet, il est primordial d'identifier précisément le lot à traiter. Pour cela, la collection doit être triée et indexée.

Dans un second temps, un travail de tri semble nécessaire afin de séparer les articles (reprints) des thèses. En effet, les articles n'ont peu de valeur pour la bibliothèque étant donné qu'ils peuvent être consultés auprès des éditeurs de revue. Les thèses par contre peuvent avoir une certaine valeur selon leur degré de rareté.

La numérisation de masse ne semble pas être une solution dans le cas présent à moins que ce procédé soit peu coûteux et que l'extraction des métadonnées soit semi automatique. Les thèses sont très structurées dans leur présentation ce qui facilite le catalogage automatique.

En outre, il sera important de vérifier la disponibilité des thèses sous format électronique via le projet NumDam (Numérisation des documents anciens de mathématiques) Source : http://www.numdam.org/

Ce programme de numérisation est une initiative conjointe du Ministère français de la recherche et du CNRS. Le programme NUMDAM de la Cellule MathDoc soutient les revues de mathématiques en rendant leurs archives visibles sur la Toile et facilement accessibles. Ce soutien consiste donc en la mise en place d'un libre accès aux données bibliographiques et au texte intégral des articles parus dans les revues. Pour chaque revue concernée, l'ensemble des volumes publiés jusqu'en l'an 2000 ont été convertis au format numérique. Cette vérification permettra de cibler les pièces rares ou uniques de la collection. De plus, celle-ci visera à éviter de renumériser ce qui a été effectué ailleurs.

Par ailleurs, l'âge des documents étant variable, il sera nécessaire de vérifier pour chaque pièce, les droits de diffusion conformément à la loi sur le droit d'auteur (LDA).

Après avoir consulté un échantillon du fonds, il s'avère qu'un nombre non négligeable de documents se dégradent. Il sera donc nécessaire de restaurer une partie de la collection avant de la soumettre au traitement des scanners. Il n'existe pas d'atelier de reliure à l'EPFL. Cependant, une documentaliste a pour charge de réparer certains documents endommagés. Pour un travail d'envergure, il sera plus judicieux de se tourner vers un prestataire externe pour la prise en charge de documents anciens et fragiles.

5.1.1.3 Coûts de la numérisation

Ne connaissant pas le nombre exact de document à écarter, il est également difficile d'établir une estimation. Pour cette collection, le coût de numérisation peut être évalué comme suit :

Ces documents nécessitent le traitement suivant : résolution de 300 dpi, niveau de gris, format PDF. La résolution de 300 dpi est un standard de nos jours. Il n'est pas nécessaire de monter plus haut en résolution pour ce type de document. L'intégrité du document est respectée. Par ailleurs, la gestion des couleurs ne semble pas être d'actualité car l'échantillon de documents consulté ne contient aucune image en

couleur. Ce sont, pour l'essentiel, des schémas et tableaux. La numérisation prenant en charge le niveau de gris, s'applique pour les schémas et croquis en demi-teintes ainsi que pour les documents faiblement contrastés. La profondeur d'acquisition noir/blanc peut également convenir à ce type de document. Cependant, dans le cas présent, il est préférable d'éviter de prendre des risques compte tenu de l'impossibilité de vérifier l'état de chaque document. Le format PDF s'adapte parfaitement à ce traitement. De plus, la majorité des documents déposés dans Infoscience sont disponibles dans ce format.

Voici mon estimation chiffrée (prix en CHF):

Scanning : 0.30 ct par page.

Séparation des pages et détourage : 0.05 ct par page.

Traitement d'image : 0.05 ct par page.

(OCR (si nécessaires) : 0.10 ct la page.

Agrégation et nommage fichiers : 1.50 CHF par document.

Prise en charge par document (livre, classeur, boîte d'archives) : 5.00 CHF.

Total : si nous nous basons sur une moyenne de 100 pages par document, nous atteignons la somme de 300'000 CHF (6000 x 0,5 x 100). Nous ajoutons 39'000 CHF pour l'agrégation et le nommage des fichiers ainsi que la prise en charge des documents. Ceci nous donne un total de 339'000 CHF.

5.1.2 Documents de la Communauté d'études pour l'aménagement du territoire - CEAT (ENAC)

Propriétaire de la collection: M. Schuler, professeur titulaire de la Communauté d'études.

Nombre de documents: 4703 recensés avant le transfert à la bibliothèque de l'EPFL. 2795 recensés après le transfert.

Localisation : EPFL, Bâtiment BP, Station 16, 1015 Lausanne (VD).

Utilisateurs/usagers: chercheurs, Offices cantonaux, étudiants, grand public.

Droits: EPFL / éditeurs commerciaux.

5.1.2.1 Présentation de l'institution

La CEAT est une plateforme de coordination pour la recherche au niveau romand, rattachée à la faculté de l'environnement naturel, architectural et construit (ENAC). La Communauté d'études a été créée par les cantons romands en 1970 dans le but de regrouper les professionnels de l'aménagement du territoire. Cette institution scientifique travaille sur des mandats émanant des Villes, des Cantons et de la Confédération dans le domaine de la recherche. En parallèle à cela, la CEAT est active dans l'enseignement et dispense des cours à l'EPFL.

L'organe de direction de la CEAT, le Conseil, est composé de personnalités scientifiques et de praticiens de l'aménagement du territoire issus des cantons et des Hautes écoles de Suisse occidentale, de la Conférence universitaire de Suisse occidentale (CUSO) ainsi que de l'Office fédéral du développement territorial (ODT-ARE). Il est actuellement présidé par Jean-Michel Cina, conseiller d'État du Canton du Valais. L'organe exécutif est le secrétariat général, formé d'une équipe pluridisciplinaire dirigée par Martin Schuler.

5.1.2.2 Présentation de la collection

Elisabeth Becker, bibliothécaire, est en charge de la gestion documentaire au sein de la CEAT.

Le développement de la collection a débuté dans les années 70. Les références sont cataloguées dans le logiciel de gestion de bibliothèque Bibliomaker. Le fonds documentaire comporte essentiellement de la littérature grise tel que des rapports, publiés ou non, d'autres institutions, ainsi que des articles et des tirés à part. En outre, des plans directeurs de grand format sont conservés dans des classeurs à rubriques. Une solution de centralisation de la masse documentaire permettrait un renforcement de la collaboration entre chaque entité.

La gestion du prêt cible les demandes internes et externes à l'EPFL. Les documents peuvent être envoyés directement à l'adresse des membres de la communauté CEAT sans qu'ils aient besoin de se déplacer.

Dans le cas présent, une numérisation des ouvrages de référence impliquera une négociation avec les ayants droit. En ce qui concerne la production scientifique et administrative, la gestion des droits revient à la CEAT ou tombe dans le domaine public.

Actuellement, il est très compliqué de regrouper les différentes collections car le travail de tri est trop important, dû principalement à la dispersion géographique des documents. De plus, sera-t-il nécessaire de tout numériser dans ce fonds? Etant donné la pluralité des acteurs de la communauté, il sera primordial de prendre connaissance des autres projets de numérisation en préparation sur cette thématique. Pour ce fonds, il ne semble pas nécessaire, hormis les plans directeurs, de conserver les documents papier parallèlement au numérique. Il sera cependant nécessaire de délimiter plus précisément la documentation en vue de la numérisation.

Les usagers sont principalement des chercheurs de la CEAT, les Offices cantonaux, les étudiants universitaires et éventuellement le grand public.

Tableau 2

Estimation des documents à prendre en considération lors de la numérisation (chiffres provisoires au 23.06.2011)

Cotes	CEAT	Doublons	Ecartés	A considérer
B - monographies A4	772	?	?	772
C - documentation A5	276	?	?	276
CD - CD-ROM	33	?	33	0
CS - séries A5	17	?	?	17
D - documentation A4	1155	?	?	1155
DS - séries A4	189	121	58	10
OT - Outil de travail	4	0	4	0
X - publications CEAT	239	?	0	239
XC - cours CEAT	28	0	28	0
Z - plans directeurs	82	?	19	63
Total	2795	121	142	2532

A – monographie A5 : ressources documentaires ayant servi aux travaux de recherche.

B – monographies A4 : ressources documentaires ayant servi aux travaux de recherche. Documents à la reliure rigide.

C – documentation A5 : littérature grise (rapports).

CD – CD-ROM : actes de colloques, congrès, statistiques etc.

CS – séries A5 : rapports publiés sous forme de série.

D – documentation A4 : ressources documentaires ayant servi aux travaux de recherche. Documents à la reliure souple.

DS – séries A4 : rapports publiés sous forme de série.

OT – outil de travail : ouvrages mis temporairement à la disposition des chercheurs.

X – publications CEAT.

XC – cours CEAT : polycopiés, transparents.

Z – plans directeurs.

Consultation : quotidienne pour la littérature grise et les périodiques. Pour les monographies (cotes A et B) stockées dans les compactus, la consultation est rare.

5.1.2.3 Recommandations pour la numérisation

La gestion active de la bibliothèque de la CEAT, m'a permis d'identifier rapidement les types de document présents dans les locaux. De concert avec Chantal Blanc, nous avons défini différents critères, tels que la valeur informationnelle, le taux de consultation, l'état physique du document, etc. Ces critères permettront de sélectionner les documents pertinents pour la numérisation.

Nous estimons que le nombre de documents sera suffisamment important pour soumettre le travail de numérisation à un prestataire externe. Cependant, les ressources documentaires ayant contribué aux travaux de recherche sont soumis au droit d'auteur. Il sera donc nécessaire de négocier avec les ayants droit si la décision est prise de les numériser. En ce qui concerne les documents publiés par la CEAT (rapports, cours, etc.) la propriété en revient à la CEAT.

A l'instar de la collection de tirés à part de mathématique, une numérisation de masse ne semble pas être un procédé judicieux. De nombreux documents (environ 2000) ont été déposés dans un dépôt d'archive. Ceux-ci ne sont pour ainsi dire plus consultés et en attente de traitement.

5.1.2.4 Coût de la numérisation

Pour le calcul des coûts de numérisation je me suis basé sur les données transmises par M. Rod, de l'entreprise 4DigitalBooks. Etant donné que la sélection des documents sera effectuée après le dépôt de mon travail, j'ai considéré comme base de calcul les chiffres provisoires (colonne "à considérer") du tableau 2.

Les plans directeurs nécessitent le traitement suivant: scan à 300 dpi, gestion des couleurs, format PDF.

Scanning format A1: 10 CHF par document

Scanning A0: 20 CHF par document

Détourage: 2 CHF par document

Traitement d'image: 2 CHF par document

Total: pour les 63 plans directeurs (50% A1 et 50% A0) le montant s'élève à 1197 CHF (756 CHF format A0 + 441 CHF format A1)

La littérature grise (publications CEAT, cours, etc.) nécessite le traitement suivant: scan 300 dpi, niveau de gris, PDF. Cette littérature étant consultée couramment, mes arguments sont les même que pour les thèses de mathématique en ce qui concerne la profondeur d'acquisition et la résolution.

La documentation A5 et la série A5 :

Scanning : 0.20 par page

Détourage : 0.05 par page

Traitement d'image : 0.05 par page

OCR (si nécessaires) 0.10 la page

Total : pour la documentation A5 et la série A5 (293 documents), le montant s'élève à 23'440 CHF si nous prenons en compte une moyenne de 200 pages par document.

Nous y ajoutons les frais d'agrégation et le nommage des fichiers ainsi que la prise en charge des documents qui s'élève à 1904.5 CHF (6.5 CHF x 293) au total et nous obtenons la somme de 25'344.5 CHF pour le traitement du lot.

Pour la série A4 et les publications CEAT (249 documents) :

Scanning : 0.30 par page

Séparation des pages et détourage 0.05 par page

Traitement d'image : 0.05 par page

OCR (si nécessaires) 0.10 la page

Total : pour la série A4 et les publications CEAT, le montant s'élève à 24'900 CHF si nous prenons en compte une moyenne de 200 pages par documents.

Nous y ajoutons les frais d'agrégation et le nommage des fichiers ainsi que la prise en charge des documents qui s'élève à 1618.5 CHF (6.5 CHF x 249) au total et nous obtenons la somme de 26518.5 CHF pour le traitement du lot.

5.1.3 Documentation du laboratoire de construction en béton – IBETON (ENAC)

Propriétaire de la collection: Aurelio Muttoni et son adjoint Olivier Burdet.

Nombre de documents: 5649 références dans Barbie.

Localisation: EPFL, Bâtiment GC B2, Station 18, 1015 Lausanne (VD).

Utilisateurs/usagers: chercheurs, étudiants, professeurs.

Droits: EPFL / éditeurs commerciaux.

5.1.3.1 Présentation du laboratoire

Le laboratoire de construction en béton est rattaché à la faculté de l'environnement naturel, architectural et construit (ENAC). Le laboratoire est actif dans plusieurs domaines de recherche tels que le poinçonnement des dalles en béton, les structures en béton fibrés à ultra-hautes performances et autres méthodes de champs de contraintes.

5.1.3.2 Présentation de la collection

L'IBETON possède un petit centre de documentation qui est tombé en désuétude lors du départ du bibliothécaire. Selon M. Burdet, remettre en activité ce service nécessite un investissement de 30'000 CHF. Cette somme est disproportionnée par rapport aux besoins réels. Chez l'IBETON, le document est un produit de consommation, où les intéressés procèdent à l'achat direct. L'intérêt se porte sur les proceedings et les livres contenant des méthodes de calcul. Ces derniers permettent une analyse simple et efficace contrairement aux logiciels électroniques qui demande de bonnes connaissances selon M.Burdet. Les travaux du laboratoire se basent généralement sur des articles scientifiques et des normes de construction suisses, européennes et mondiales. Pour celles-ci, un accès électronique direct et la possibilité d'achat à la pièce est souhaitable. A l'heure actuelle, seule la bibliothèque de l'EPFL dispose d'un poste unique où le recueil des normes SIA (**S**ociété suisse des **I**ngénieurs et des

Architectes) est consultable. La bibliothèque a les droits d'impression mais n'est pas habilitée à diffuser les documents de manière électronique.

Le centre de documentation possède des ouvrages de référence mais il n'existe pas de catalogue mis à jour. L'état des documents est bon. Cependant, M. Burdet ne voit pas l'intérêt de dépenser une somme importante pour la numérisation de toute la collection car une majorité des documents n'est plus utilisée. Les laboratoires ne se préoccupent pas de la mémoire patrimoniale. Certains documents datent des années 70 (rares et important). Ils ont une réelle valeur pour comprendre les constructions de l'époque. Par ailleurs, il est nécessaire de conserver les normes. Sous une forme numérique, ces normes seraient beaucoup plus accessibles. Actuellement, seuls les membres SIA du laboratoire bénéficie d'un accès électronique direct à celles-ci. M. Burdet désire installer 4 licences supplémentaires à la base de données bibliographique E-reader pour faciliter l'accès aux normes SIA. Selon M. Burdet, 5 documents sont vraiment utiles dans leur centre de documentation.

En ce qui concerne les publications, la production scientifique est transmise à M. Favre (bibliothécaire) sous forme de fichiers XML pour le dépôt dans Infoscience après avoir été inscrit préalablement dans Barbie (**B**ase d'**AR**ticles **BI**bliographiqu**E**s), la base de données bibliographique du laboratoire. Tout est transféré dans Infoscience (travaux des chercheurs de L'EPFL ou ayant travaillé pour le compte de l'école) sauf les polycopiés et les contrats d'exclusivité. Les articles sont disponibles en texte intégral (PDF) pour les personnes autorisées. La production antérieure à 1990 n'est pas en ligne. Le scan s'effectue au compte goutte selon les besoins. La secrétaire a débuté ce travail de numérisation des anciennes publications dans le but d'un accès facilité. Pour cela, une imprimante multifonction (Brother) se charge de numériser les documents à une résolution de 150 dpi. Il est possible de monter la résolution à 300 dpi le cas échéant. La gestion des niveaux de gris doit être satisfaisante pour retranscrire correctement les nombreux croquis et schémas. Dans une moindre mesure, selon M. Burdet, le scan électronique de Nebis répond également à leurs besoins en terme de qualité de numérisation. Lorsque nous avons fait remarquer à M. Burdet que la bibliothèque disposait d'un scanner à livre, il s'est montré très intéressé. Cet outil permettra d'accélérer le processus de numérisation pour les documents reliés.

La masse documentaire se compose d'articles scientifiques, d'articles de conférence, de diapositives (scannées et très demandées). Les éditeurs ont les droits sur certains documents. Force est de constater que de nombreux articles n'ont pas d'intérêt à être numérisé. Depuis 2000, il n'y a plus d'acquisition d'ouvrage dans le centre de

documentation. Les livres commandés vont directement dans les bureaux des collaborateurs (commandes personnelles).

De surcroît, il existe un réel problème de récupération de l'information. Les titres de certains ouvrages ne sont pas toujours en relation directe avec le contenu, ce qui crée une certaine confusion lors de la recherche. Par exemple, la collection ITC pour les abris de protection. Un gros travail sera sans doute nécessaire pour entrer les métadonnées lors de l'indexation des documents.

En outre, il existe un réel problème d'archivage. Depuis que l'OFrou (Office fédérale des routes) gère les autoroutes, les archives cantonales ont migrées chez eux. Cela rend leur consultation compliquée. Dorénavant il est nécessaire de transiter par leur service documentaire avant de pouvoir consulter le document souhaité, ce qui entraine de nouveaux délais.

Le projet de numérisation mérite donc d'être lancé selon M. Burdet, cependant, la problématique liée à la pérennité des données l'inquiète tout particulièrement. Pour éviter tout incident, il garde une copie privée de toutes les publications qu'il envoie à M. Favre, destinées à alimenter les archives institutionnelles de l'EPFL.

5.1.3.3 Base d'articles bibliographiques (Barbie)

Créé en 1994 par des étudiants, Barbie est une base de données d'articles scientifiques dans le domaine du génie civil. Cet outil est principalement destiné aux collaborateurs de la faculté Enac. Depuis peu cette base de données est utilisée uniquement par les collaborateurs de l'IBETON. Barbie offre un stockage standardisé des références bibliographiques, une gestion facilitée de listes de références et la possibilité d'exporter les données pour une utilisation dans le cadre de publications. De plus, les articles sont disponibles en texte intégral (PDF). La base de données recense 5649 références bibliographiques à l'heure actuelle.

L'accès à la base est disponible sur Internet avec un accès possible pour la consultation à de nombreux chercheurs en dehors de l'EPFL.

5.1.3.4 Remarques

A la suite de mes entretiens avec M. Burdet, j'ai noté que leur principal intérêt à l'IBETON se situe au niveau d'un accès facilité aux normes SIA ainsi qu'à l'acquisition de nouveaux périodiques spécialisés. Le travail de numérisation serait pour eux un atout supplémentaire mais ne constitue pas une réelle priorité.

Types de documents : Monographies, proceedings, normes, catalogues.

Formats des documents : A4 et A5.

Nombres de documents : environ 800 documents.

Etat de la collection : bon.

Consultation : quotidienne.

Documents à numériser : environ 50.

Tableau récapitulatif des documents présents à l'IBETON

Types de document	Quantité
Monographies	200
Périodiques	20
Normes	?
Catalogues	?
Proceedings	500 à 600

Voici un échantillon de documents présents au centre de documentation du laboratoire qui ont un intérêt à être numérisé selon M.Burdet :

- Les classeurs VSL (expert des systèmes de précontrainte et de haubanage) contiennent les catalogues du fournisseur de système de précontraintes. Ces documents sont orientés certes, mais sont utiles pour les recherches des doctorants.
- Einflussfelder elastischer Platten /Adolf Pucher. Adolf Pucher était professeur en génie civile et basait son cours sur ses propres ouvrages. Cette série de monographies est utile aux étudiants selon M. Burdet.
- Vorlesungen über Massivbau. Editeur Springer.

Ces références représentent environ 50 documents.

5.1.3.5 Recommandations pour la numérisation

La masse de document n'est pas suffisante selon moi pour mettre en place un appel d'offre aux prestataires. Ici, il ne s'agit pas de numériser pour numériser mais de sélectionner les documents de manière ciblée. Par ailleurs, en l'état actuel des choses, les outils de numérisation disponibles à la bibliothèque de l'EPFL sont adaptés aux besoins de numérisation du laboratoire. Nous pouvons imaginer qu'un étudiant soit mis à contribution durant la trêve estivale pour effectuer le scan des documents. Les économies effectuées pourront être réallouée à un autre lot de document.

5.1.4 Documentation du laboratoire de la construction métallique – ICOM (ENAC)

Directeur: Prof. Dr. Jean-Paul Lebet.

Propriétaire de la collection: Alain Nussbaumer, professeur titulaire et responsable du domaine de recherche « fatigue et rupture des structures en acier ».

Nombre de documents: environ 3960 documents recensés dans leur base de données.

Localisation: EPFL, Bâtiment GC B3, Station 18, 1015 Lausanne (VD).

Utilisateurs/usagers: étudiants, professeurs, chercheurs, ingénieurs de la pratique.

Droits: EPFL / éditeurs commerciaux.

5.1.4.1 Présentation du laboratoire

Fondé en 1969, l'ICOM est le plus ancien laboratoire de l'EPFL. Il est rattaché à la faculté de l'environnement naturel, architectural et construit (ENAC). Dans le secteur de l'enseignement, le laboratoire enseigne la conception et le dimensionnement des structures métalliques. Parallèlement à cet activité, l'ICOM mène une recherche théorique et appliquée dans ce secteur et propose ses services dans le cadre de travaux pour des tiers (industries, associations scientifiques et professionnelles, etc.).

5.1.4.2 Présentation de la collection

L'ICOM possède une bibliothèque active, gérée par Claudio Leonardi, assistant technique travaillant à 40% dans cette unité documentaire. La base de données de la bibliothèque (BiblBD), créée en l'an 2000, recense 3960 références bibliographiques et propose des documents en texte intégral au format PDF. Ces références correspondent à différents types de documents : monographies (1294), périodiques (145 titres), dictionnaires, rapports (617), diapositives (environ 8000 pièces), proceedings (400), etc. Les rapports d'expertise, souvent confidentiels, sont gérés par le secrétariat. Le logiciel FileMaker Pro 6 sert de catalogue aux usagers. Chaque usager possède sont propre identifiant pour se connecter à la base de données. La recherche s'effectue à l'aide de nombreux champs (auteurs, titre, mots clés, langues, années d'édition, éditeur, n° ISBN, etc.). Le catalogage des documents est pris en charge par la bibliothèque et les doctorants, selon une relation de confiance.

Le recensement des rapports ou publications ICOM prend fin en 2005. A partir de cette date, les documents sont référencés dans Infoscience. Parmi ces publications, une centaine sont réellement intéressantes et mériteraient un traitement numérique selon

M. Nussbaumer. Par publications intéressantes, nous entendons les documents de référence sur les pratiques de construction qui sont toujours d'actualité malgré les évolutions technologiques. Les publications antérieures à 2000 sont présentes uniquement sous forme papier. Les documents les plus demandés ont déjà été scannés au compte goutte à l'aide d'imprimantes multifonctions.

Il existe également des photographies sous forme de diapositive stockées dans une armoire prévue à cet effet ou dans des classeurs. Celles-ci sont peu demandées par les collaborateurs mais possèdent une valeur historique indéniable pour certaines. Classées par thématiques dans des classeurs, les diapositives n'ont pas été indexées. Ces images étaient utilisées par le passé pour l'enseignement. Elles illustrent les exemples de la pratique tels que l'ossature des bâtiments de l'EPFL, les ponts construits de part le monde, etc. Cependant, de nombreux clichés n'ont aucune valeur scientifique et doivent être jetée selon M. Nussbaumer. Un travail de tri sera donc nécessaire. Pour cela, il serait judicieux de contacter les anciens collaborateurs et professeurs ICOM ayant travaillé avec ces images. Ces personnes connaissent mieux que quiconque les circonstances dans lesquels les clichés ont été capturés. Après les avoir numérisées, les diapositives seront stockées probablement sur le serveur de la faculté ENAC.

A l'instar du laboratoire IBETON, l'ICOM travaille avec les normes SIA. La collection des normes SIA comprend des normes, des règlements, des directives, des recommandations et des cahiers techniques. Ces normes sont disponibles en version papier dans de grands classeurs blancs ou en format électronique avec abonnement. Les normes au format électronique sont classées dans une base de données à part au sein de l'ICOM. Les postes de consultation sont limités et les étudiants doivent se rendre à la bibliothèque de l'EPFL pour consulter et/ou imprimer celles-ci. L'acquisition de plusieurs postes E-reader pour la consultation des normes serait une solution appréciée au sein du laboratoire.

Tout au long de son existence, les collaborateurs de l'ICOM ont énormément publiés tout en appliquant une méthode de conservation efficace. Cependant, les locaux ne pourront plus contenir cette masse de documents éternellement. Le personnel a commencé à désherber la collection afin d'éviter la saturation.

Par ailleurs, M. Leonardi a débuté la création d'un programme de gestion des figures présentes dans les publications ICOM. Ce programme permet de retrouver et d'organiser des croquis et autres tableaux afin de les réutiliser pour de futures recherches ou pour l'enseignement. Cependant, le projet n'a pas pu être mené à son

terme et le programme est tombé dans l'oubli. Force est de constater qu'un tel programme peut être d'une grande utilité pour la communauté ENAC.

5.1.4.3 Les besoins des usagers

Les collaborateurs désirent retrouver facilement les bases, les sources, les documents d'origine des différents projets. Comment ont été fait l'essai, les mesures? Force est de constater que les articles ne le précisent pas systématiquement. Les sources sont parfois au sein du laboratoire mais peuvent être aussi à l'extérieur. Cela peut poser des problèmes pour la consultation.

Les usagers de la collection sont principalement des doctorants, des chercheurs et des ingénieurs de la pratique.

Tableau récapitulatif des documents présents à l'ICOM

Types de documents	Formats	Quantité
Monographies	A4	1294
Périodiques	A4	145 titres
Publications ICOM	A4	617
Thèses	A4 ; A5	?
Diapositives	diapositives sur film standard	environ 8000
Proceedings	CD-ROM et format papier A5	400
Rapports d'expertise	?	?

5.1.4.4 Recommandations de numérisation

A l'instar de l'IBETON, le laboratoire effectue au compte goutte une numérisation des documents consultés. Cependant, une numérisation de masse de la collection semble être disproportionnée au vu des réels besoins des usagers. Il sera, ici, plus utile de sélectionner consciencieusement les documents prêts à être scannés. Selon M. Nussbaumer, une partie des diapositives et une centaine de publications ICOM méritent d'être numérisé. Il est à noter que les publications ICOM peuvent être numérisées aisément avec les outils présents à la bibliothèque de l'EPFL. Le nombre de documents et leur format ne sont pas de réels obstacles à la numérisation interne, ce qui n'est pas le cas pour les diapositives.

5.1.4.5 Coûts de la numérisation

A ma connaissance, il n'existe pas de machine capable de numériser des diapositives sur film au sein de l'EPFL. Ce traitement devra donc être délégué à un prestataire externe. Le nombre total d'images retenues pour le projet de numérisation est

inconnue pour l'heure. Mes estimations se basent sur un lot de 4000 diapositives, soit la moitié de la collection. Les tarifs sont indiqués en francs suisses.

Les diapositives nécessitent le traitement suivant : 300 dpi, gestion des couleurs, jpeg.

Scanning : 0,30 par diapositive

Réajustement des niveaux et des couleurs : 0,10

Restauration des images usées : 0,80

Total : pour les 4000 diapositives, le montant s'élève à 1600 CHF sans compter les frais supplémentaires engendrés par les images usées.

Les publications ICOM nécessitent le traitement suivant : 300 dpi, niveaux de gris, PDF.

Scanning : 0.30 par page

Séparation des pages et détourage 0.05 par page

Traitement d'image : 0.05 par page

(OCR (si nécessaires) 0.10 la page

Total : pour les 100 publications, le montant s'élève à 50 CHF. Nous y ajoutons les frais d'agrégation et le nommage des fichiers ainsi que la prise en charge des documents qui s'élève à 650 CHF (6.5 CHF x 100) au total et nous obtenons la somme de 700 CHF pour le traitement du lot.

5.1.5 Documentation du laboratoire des machines hydrauliques – LMH

Propriétaire: François Avellan, professeur en ingénierie hydraulique.

Nombre de documents: 4000.

Localisation: Avenue de Cour 33 Bis, 1007 Lausanne (VD).

Utilisateurs/usagers: scientifiques et doctorants.

Droits: Editeurs commerciaux / EPFL.

5.1.5.1 Présentation du laboratoire

Le laboratoire des machines hydrauliques (LMH) à Lausanne se situe hors du campus de l'EPFL. Le LMH est rattaché à la faculté des sciences et techniques de l'ingénieur (STI). Leur pôle de compétence se concentre sur les installations hydroélectriques

telles que les turbines hydrauliques. Le laboratoire effectue des expertises de rendement de différentes installations à travers le monde. Au sein de leurs locaux, 3 machines sont utilisées pour mener des simulations et autres bancs d'essai.

Le laboratoire a été créé en 1969 par le Professeur Bovet. Au fil de son existence, le LMH a accumulé un savoir-faire sur toutes les formes de turbines et aménagements hydrauliques. Parallèlement à son activité première, le laboratoire dispense des cours aux doctorants spécialisés dans la mécanique des fluides.

5.1.5.2 Présentation de la collection

Le LMH possède une bibliothèque contenant divers ouvrages et périodiques dans le domaine des mathématiques, de la physique, de l'informatique etc. D'anciennes normes CEI (communauté électrotechnique internationale) et ISO (organisation internationale de normalisation) parsèment les étalages. Malheureusement, la bibliothèque n'est plus tenue à jour. La gestion documentaire de cette petite entité s'est éteinte en 2006 lorsqu'il a été décidé de ne plus repourvoir le poste de bibliothécaire. Dès lors, les scientifiques ont repris le flambeau en essayant de maintenir un semblant d'ordre dans les locaux. Cependant, sans service de prêt, de nombreux ouvrages ont disparu des rayons sans espoir d'y retrouver leur place. Dorénavant, les commandes de documents s'effectuent à titre individuel et se retrouvent directement dans les bureaux des intéressés. Les demandes documentaires sont très faibles à ce jour. La bibliothèque est vouée à disparaître et sera probablement transformée en dépôt d'archives.

Les publications du laboratoire se composent de rapports d'expertise, souvent confidentiels, d'articles scientifiques et de proceedings. Les doctorants publient leur thèse et déposent un stock de documents au LMH afin que celle-ci soit diffusée aux intéressés. Il existe des thèses datant de 10 ans et plus qui ne sont pas disponibles en format électronique. Ceci est à vérifier car je n'ai pas pu obtenir les références de ces thèses.

Un travail de numérisation des documents présents à la bibliothèque ne semble pas être pertinent de prime abord. Une multitude d'ouvrages sont devenus désuets au fil de temps et ne présentent plus d'intérêt aux yeux de la communauté scientifique. La plupart des ouvrages consultés sont référencés au sein de la bibliothèque du Rolex Learning Center. Cela rend d'autant plus difficile de maintenir en activité ce centre de documentation. Par ailleurs, les doctorants utilisent plus facilement Infoscience ou d'autres outils d'information pour la recherche documentaire.

Cependant, la numérisation permettrait de diffuser plus aisément les thèses des doctorants sans devoir gérer les stocks de documents. Comme mentionné plus haut, une dizaine de thèses ne sont pas disponible sous forme électronique. Selon M. Farhat, le travail de numérisation doit se concentrer sur ces documents spécifiques. Le reste de la collection n'appartient pas au laboratoire, il sera donc compliqué d'obtenir les droits de numérisation. En outre, l'intérêt devient faible lorsqu'il n'existe pas de demande ni de besoin documentaire au sein du centre de documentation. Numériser la totalité de la collection serait une perte de temps et de ressources financières selon les collaborateurs du LMH.

Tableau récapitulatif des documents présents au LMH

Types de document	Quantité
Monographies	environ 4000
Périodiques	environ 10 titres
Thèses	10 non numérisées

La bibliothécaire en charge de la collection étant partie à la retraite, je n'ai pas réussi à récolter suffisamment d'informations pour établir un inventaire précis de la collection. Les scientifiques sur place n'ont qu'une connaissance limitée de l'état de la collection. Néanmoins, j'ai mis la main sur un fichier XML regroupant les nombreuses entrées du catalogue Bibliomaker. Le fichier n'étant pas « bien formé » ni rattaché à une DTD, je n'ai pas pu trier les notices pour en tirer les informations nécessaires.

5.1.6 Gestion des archives courantes au Service académique – SAC

Pour conclure l'état des lieux des entités étudiées, voici l'exemple du service académique et de sa gestion des dossiers d'étudiants.

Le service académique est le service administratif responsable de la gestion et de la conservation des dossiers des étudiants pour les études de Bachelor, Master, Doctorat et formation continue. Il est composé d'une quinzaine de spécialistes avec des connaissances ciblées selon les types d'études et les questions soulevées.

Le service s'occupe principalement des opérations d'admission, d'inscription selon les règlements de l'École, ainsi que l'élaboration des horaires des cours et des examens. En outre le SAC offre de nombreuses prestations pour les étudiants et les enseignants.

A la base, le projet GED répond à des besoins similaires formulés par différents services. Avant la GED, les dossiers d'étudiants se trouvaient soit sous format papier,

soit sous format électronique. Pour le format papier, la personne désirant consulter un dossier devait se déplacer pour chercher le dit dossier. Le format électronique concernait les dossiers archivés, qui avaient été numérisés et stockés dans un système propriétaire CANOFILE. Les cassettes ne pouvaient être lues que sur un appareil bien précis qui se trouvait dans un local spécifique, distant des bureaux où travaillaient les personnes amenées à traiter ces dossiers.

Rapidement, il est devenu nécessaire de centraliser l'information pour en faciliter l'accès via une interface web commune. De plus, un accès via Internet et une application web pour simplifier les questions de maintenance et d'administration (cela se passe sur le serveur et non sur les postes utilisateurs) a été mis en place. Il était également nécessaire de sécurisé les archives papier et électroniques afin d'optimiser leur conservation physique. En effet, les dossiers papier des étudiants en cours d'étude pouvaient disparaître en cas d'incendie.

Dorénavant, l'accès aux documents au travers de IS-Académia rend la GED transparente pour les utilisateurs finaux, qui n'ont pas à apprendre à utiliser une nouvelle application.

Au sein du SAC, les documents se présentent en série, de manière structurée pour les dossiers d'étudiants et d'enseignants. C'est donc un fonds homogène ou se côtoient les bulletins de notes, les lettres d'admission, les formulaires de candidature etc.

La mise en place du projet est due à la volonté de passer d'une gestion des dossiers archivés à une gestion des dossiers courants. De plus les problèmes de pérennité du système d'archivage existant (CANOFILE) n'a fait que confirmer cette décision.

La solution retenue répond donc aux besoins spécifiques des différents services (différents projets GED de l'EPFL) en prenant compte des disparités budgétaires. Le service académique espère donc récupérer des documents administratifs, en facilitant leur prise en charge, qui pourraient rester dans les secrétariats des laboratoires (dans le cas de doctorants).

Pour le SAC, la principale difficulté réside dans le fait d'interfacer Alfresco et IS-Académia et faire qu'ils échangent des données de manière intelligente.

Afin d'alimenter le système, les dossiers courants sont soumis au processus de numérisation. Les dossiers d'archives étaient déjà numérisés, car ils faisaient partie de CANOFILE (à part quelques exceptions dues au mauvais état du système CANOFILE). Tout le projet consistait à passer d'une gestion des dossiers d'archives à une gestion

des dossiers actifs. Il est à noter que le système prend également en compte les candidatures soumises électroniquement via IS-Académia (inscription à des doctorats qui se fait uniquement sous forme électronique). IS-Académia donne accès à tous les documents via le dossier de l'étudiant. Lorsqu'un étudiant n'a pas de dossier IS-Académia, (c'est le cas pour les dossiers des étudiants ayant étudié avant l'an 2000 approximativement), les documents de ces dossiers sont uniquement accessibles via Alfresco. De cette manière, l'accès à l'information est plus centralisé et facilité.

Les dossiers courants papier sont numérisés et indexés avant d'être intégrés dans Alfresco. Par ailleurs, un module permet l'intégration des emails importants dans la constitution des dossiers. Il est à noter qu'Alfresco donne accès aux documents tandis que IS-Académia donne un accès aux données métiers. En d'autres termes, Alfresco héberge les documents et la consultation de ceux-ci s'effectue via IS-Académia.

Le scan des documents s'effectue à l'aide d'une machine KOFAX de type feuille par feuille. Les documents numérisés sont ainsi directement intégrés dans Alfresco grâce à la reconnaissance de code barre et le classement automatique des documents. Le plan de classement comporte plusieurs niveaux. Un dossier étudiant contient toutes les personnes en formation. Pour chaque individu, des métadonnées descriptives identifient le profil. Le dossier d'un étudiant lambda se divise en plusieurs parties suivant le cursus académique suivi (Bachelor, Master, Doctorat). Cette arborescence permet une gestion des documents efficace.

En outre, chaque étudiant se voit confier un numéro SCIPER qui permet de l'identifier. Les documents relatifs à l'étudiant sont accompagnés d'un code barre permettant de faire le lien avec ce dernier.

Dès lors que le document papier est numérisé, on procède à l'évaluation de la qualité du fichier nouvellement créé afin de corriger les anomalies. Par la suite, l'indexation (saisie des métadonnées) se fait de manière manuelle. En principe l'indexation et le classement sont automatiques grâce au code barre apposé sur les documents avant leur numérisation. Le seul cas où il est nécessaire d'intervenir manuellement est lorsque pour une raison ou une autre le code barre n'a pas pu être lu. Dans ce cas le document est retenu par KOFAX et envoyé dans un « endroit spécial » où il sera traité manuellement, puis réinjecté dans le circuit usuel pour être importé dans Alfresco.

5.2 Les besoins et les attentes des différents laboratoires

De nos jours, les avancées technologiques dans le domaine de la recherche documentaire ont profondément modifié nos comportements. Les chercheurs et le

grand public s'attendent à pouvoir disposer de tout sur le web. Ceci de manière immédiate, permanente et de préférence gratuitement. Les bibliothèques numériques ont pour objectif de diffuser plus largement l'information dans de nombreux formats au travers de collaborations accrues entre les différents acteurs en place (bibliothèques, centres de documentation, archives etc.).

Les usagers des centres de documentation des laboratoires se répartissent en deux groupes distincts : les scientifiques et les doctorants.

Afin de prendre connaissance des besoins et attentes de ces communautés, je me suis rendu dans leurs locaux pour un entretien avec les différents responsables. Je leur ai soumis une liste de question afin de cerner plus précisément les besoins et attentes ainsi que les enjeux d'un tel projet de numérisation. Force est de constater que la recherche documentaire n'est pas facilitée par l'absence de véritable centre de documentation actif au sein de certaines entités. Les demandes des doctorants doivent parfois être traitées par les professeurs, qui disposent des versions papier lorsque les documents numériques sont indisponibles à la consultation. Par ailleurs, le manque de coopération entre les laboratoires à ce sujet agit comme un véritable frein à la recherche. Chaque entité procède selon sa propre règle interne et ne se préoccupe pas ou peu de l'aspect collaboratif.

6. Les plateformes d'archivage

6.1 Introduction

Lorsque l'on projette de créer un logiciel ou un site web, il y a plusieurs considérations d'usage à prendre en compte. Il est important de connaître les spécificités des usagers, leur identité et leurs besoins. Il est à noter que l'utilisabilité d'un tel outil dépend grandement de l'interaction entre l'utilisateur et l'ordinateur. Cette interaction doit être efficace, facile et doit satisfaire les besoins informationnels de l'individu.

Les objectifs et les ambitions d'un site web, portail ou autre sont défini par les besoins des utilisateurs. Ces critères définissent les fonctions qui devront être intégrées au système.

6.2 Les archives institutionnelles

Une archive institutionnelle est l'archive d'une institution regroupant l'ensemble de sa production (de recherche, patrimoniale, pédagogique, administrative...) dans des espaces privatifs ou ouverts, comme le fait le CERN.

En d'autres termes, une archive institutionnelle relève d'une institution (université, grande école, organisme de recherche, association professionnelle) et a pour objectif de contenir, valoriser et conserver l'ensemble de la production scientifique de celle-ci.

6.3 Les archives ouvertes et le protocole OAI-PMH

6.3.1 Les archives ouvertes

Le terme archive ouverte désigne un réservoir où sont déposées des données issues de la recherche scientifique et de l'enseignement et dont l'accès se veut ouvert c'est-à-dire sans barrière. Cette ouverture est rendue possible par l'utilisation de protocoles communs qui facilitent l'accessibilité de contenus provenant de plusieurs entrepôts maintenus par différents fournisseurs de données. Pour être considérée comme ouvertes, les archives doivent se soumettre au protocole OAI-PMH.

Les archives ouvertes reposent sur un principe relativement simple : le dépôt, sous forme électronique, de publications scientifiques dans des entrepôts numériques (repositories).

C'est au début des années 1990 au sein des communautés scientifiques que les premières archives ouvertes se sont développées dans l'optique de faciliter la communication scientifique.

Il existe trois types d'archives ouvertes :

- Les archives disciplinaires
- Les archives centrales
- Les archives institutionnelles

Les archives disciplinaires sont les plus anciennes. ArXiv a été lancé en 1991. Le but de ces archives est de répondre à l'amélioration de la communication scientifique. Celles-ci peuvent être centralisées en utilisant un logiciel unique. Nous pouvons citer l'exemple d'E-LIS pour les sciences de l'information et des bibliothèques.

Les archives centrales sont principalement utilisées dans le but de présenter la production scientifique d'un pays et d'en accroitre sa visibilité. En France, la plate-forme HAL sert de socle à l'archive ouverte nationale par exemple.

Les archives institutionnelles regroupent l'ensemble de la production (de recherche, patrimoniale, pédagogique, administrative...) d'une institution dans des espaces privatifs ou ouverts, comme le fait le CERN et l'EPFL par exemple.

En d'autres termes, une archive institutionnelle relève d'une institution (université, grande école, organisme de recherche, association professionnelle) et a pour objectif de contenir, valoriser et conserver l'ensemble de la production scientifique de celle-ci.

Les archives ouvertes peuvent être créées par une institution mais des ensembles plus larges peuvent se créer, fédérant l'accès à l'information au niveau national (projet DARE aux Pays-Bas, Intute en Grande-Bretagne, etc.) ou dans des archives thématiques internationales : ArXiv à la bibliothèque de l'université de Cornell (USA) pour les sciences fondamentales, PubMed Central, un service du Center for Biotechnology Information (USA) pour les sciences biologiques et de santé, etc.

6.3.2 Le protocole OAI-PMH

« Le protocole OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) est un protocole informatique fondé par l'Open Archives Initiative pour échanger des métadonnées. Il permet de constituer et de mettre à jour automatiquement des entrepôts centralisés où les métadonnées de sources diverses peuvent être interrogées simultanément. Utilisé notamment par les Archives Ouvertes et les entrepôts institutionnels, il s'est aujourd'hui largement répandu dans les institutions patrimoniales et notamment les bibliothèques. »

Source : wikipedia.org

Le protocole OAI-PMH est un moyen d'échanger sur Internet des métadonnées entre plusieurs institutions, afin de multiplier les accès aux documents numériques. Pour son bon fonctionnement, l'archive institutionnelle doit se calquer sur différents standards et protocoles lui permettant de préserver et de mettre à disposition les données.

L'OAI-PMH définit deux types d'acteurs :

- les fournisseurs de données, qui déposent leurs métadonnées sur un serveur web appelé "entrepôt",
- les fournisseurs de service qui collectent (on dit aussi "moissonnent") ces données pour les intégrer à l'index de leurs propres bibliothèques numériques.

Un même établissement peut jouer les deux rôles : diffuser ses métadonnées et collecter celles des autres.

Le fonctionnement de base du protocole OAI-PMH repose sur une communication de client à serveur. Le client envoie des requêtes au serveur en http, le serveur répond par un flux de données en XML.

6.3.3 Fedora Commons

Ce logiciel permet de gérer des archives numériques à grande échelle. Munit de nombreuses fonctionnalités et standards, Fedora Commons propose différentes utilisations : gérer des livres dans une bibliothèque ou des collections d'oeuvres d'art dans un musée, faciliter la collaboration d'étude ou de recherche dans tous les domaines dont la base de travail peut être numérisée (enregistrement sonore ou vidéo, reconstitution archéologique, manuscrit, etc...).

Fedora Commons n'est pas une infrastructure optimale pour stocker et archiver du contenu numérique. Pour palier ce manque il existe des applications qui viennent améliorer la solution informatique. Cependant, Fedora Commons s'apparente plus à une base de données orientée archivage dont l'objectif est de stocker des documents numériques.

Fedora Commons est utilisé par l'UNIL notamment.

6.3.4 Dspace

DSpace est un logiciel libre qui permet la construction d'archives électroniques ouvertes. Le logiciel est fréquemment utilisé par les Universités ou les Organismes de recherche pour stocker des collections d'articles, de thèses ou de livres. DSpace est également adapté pour l'archivage de photos, d'enregistrements sonores ou de vidéos.

Une collaboration entre le MIT et les laboratoires HP à Cambridge initiée en 2002 est à l'origine du logiciel. Les développements sont gérés depuis 2009 par Duraspace, une société américaine à but non lucratif résultant de la fusion de la DSpace Foundation et de Fedora Commons. Une communauté internationale d'utilisateurs et d'informaticiens contribue à l'évolution du logiciel.

DSpace n'est pas tout à fait équivalent à Fedora Commons. DSpace est une solution complète, avec interface utilisateur, utilisable dès l'installation. Un équivalent de DSpace est Eprints.

6.3.5 Eprints

Eprints est une solution libre et open source pour la construction d'archives institutionnelles en libre accès respectant le protocole informatique OAI-PMH. Il partage de nombreuses fonctionnalités avec les systèmes de gestion de documents mais sont usage premier se situe au niveau de la gestion des archives institutionnelles

et des périodiques scientifiques. Eprints a été développé à l'Université de Southampton sous licence GNU GPL (General Public Licence). Il est donc ouvert à tous.

6.3.6 CDS Invenio

CDS Invenio est une solution informatique offrant des outils de création et de gestion de bibliothèque numérique. C'est un logiciel libre sous licence GNU GPL. Le produit couvre tous les domaines de la gestion documentaire et respecte le protocole OAI-PMH. Le format des données bibliographiques est le MARC 21.

Développé au CERN (Organisation européenne pour la recherche nucléaire), le logiciel gère plus de 700 collections de données et plus d'un million de références bibliographiques. Par ailleurs, CDS Invenio est installé dans plusieurs institutions scientifiques dont l'EPFL.

6.4 CDS Invenio vs Dspace

Lorsque l'EPFL a dû sélectionner une solution informatique pour héberger Infoscience, deux choix se sont démarqués. CDS Invenio et Dspace ont été départagés selon différents critères. Tout d'abord, l'école s'est concentrée sur l'activité de la communauté d'utilisateur des deux concurrents. Ceci est important dans le sens qu'une communauté active répondra plus facilement aux problèmes rencontrés lors de la mise en fonction du logiciel. Sur ce point, la solution du MIT prend l'avantage avec une communauté plus grande et plus réactive sur les forums. En moyenne, 2 à 5 commentaires sont postés par jour sur les forums d'utilisateurs contre 2 à 5 par mois pour la solution du CERN.

Dans un deuxième temps, le nombre d'installations du logiciel dans diverses institutions scientifiques a été analysé. Malgré l'absence de donnée chiffrée fiable, force est de constater que Dspace a été installé plus fréquemment que la solution CDS.

L'avantage de CDS Invenio, qui s'appelait CDSWare à l'époque, consiste en la compatibilité des applications de bibliothèque avec Aleph ainsi que la localisation de l'équipe de développement en Romandie. Ces deux points revêtent une grande importance pour l'EPFL.

En outre, l'outil du CERN avait l'avantage de prendre en compte le format MARC XML contrairement à Dspace travaillant avec le format DublinCore.

De manière plus générale, CDS Invenio offrait une approche plus « user friendly » dans l'entrée et la sortie des données ainsi que dans la gestion des collections virtuelles.

Les conclusions du cabinet d'expertise FontisMedia[8] sont sans équivoque, l'outil développé par le CERN est le plus adapté au cahier des charges soumis par l'EPFL.

Voici un échantillon des arguments du cabinet d'expertise :

« For the reasons cited above, FontisMedia concludes that the CDSWare system is better suited to the specific needs of the Infoscience project than DSpace. The reasoning is summarized in the following points, organized roughly in the order of their importance:

- The XML structure (MARC based) of CDSWare is better suited to use as a bibliographic tool.
- The use for the system that is anticipated at the EPFL is closely matched to that of CDSWare at CERN.
- Development for ALEPH integration has been carried out.
- The programming of the DSpace product is based on a Java platform that might experience significant speed problems for repositories exceeding hundreds of thousands of records.
- The modules of the CDSWare are constructed in a way that allows line-command intervention for administration, testing and development.
- The CERN team is located in the region and offers bilingual support. CDSWare is proven for a system containing many hundreds of thousands of documents. A flexible toolkit that will become available as the collaboration with CERN comes online ».

6.5 Infoscience

6.5.1 Présentation

Infoscience (http://infoscience.epfl.ch/) est l'Archive institutionnelle de l'EPFL, une base de données servant à archiver et signaler les travaux et publications scientifiques (articles, papiers de conférences, proceedings, livres, chapitres de livres, posters, reports, etc.) produits par les laboratoires de l'Ecole Polytechnique fédérale de Lausanne. Cet outil permet de centraliser le patrimoine scientifique de l'EPFL et de proposer un large éventail de services aux collaborateurs afin de faciliter la collaboration et le partage. Des analyses statistiques sur les données stockées dans le dépôt sont en développement afin d'offrir aux chercheurs des outils supplémentaires. En résumé, le projet infoscience a pour but de faciliter l'accès aux ressources

[8] FontisMedia : http://www.fontismedia.com/

scientifiques produites à l'EPFL (publications, preprints, rapports de recherches, projets, thèses, travaux d'étudiants, cours, etc.)

Cette base de données est basée sur CDS-Invenio, un logiciel open source développé originellement par le Centre Européen pour la Recherche Nucléaire (CERN). Il est compatible avec le protocole de récolte de données de l'Open Archive Initiative (OAI-PMH) et utilise le format Marc XML (Machine Readable Cataloging) comme standard de description bibliographique à l'instar de la Bibliothèque du Congrès aux Etats-Unis. Ce format d'échange de données bibliographiques se composent de plusieurs champs de données normalisés par L'IFLA (**I**nternational **F**ederation of **L**ibrary **A**ssociations and Institutions).Cela favorise l'échange de données entre bibliothèques. De plus, la composante XML à l'avantage de séparer les données de leur description pour faciliter l'interopérabilité.

Infoscience est donc un outil de travail pour les chercheurs et leurs laboratoires. Ils ont la possibilité d'archiver et d'organiser leur production scientifique via cet outil.

Les archives institutionnelles sont le fondement d'Infoscience. La tendance actuelle est à l'acceptation de plus en plus de publications extérieures à l'EPFL depuis la nouvelle version mise à jour en mars 2010.

Infoscience regroupe plus de 78'000 publications EPFL dont environ 35% des documents sont disponibles en accès plein-texte. L'archive institutionnelle de l'EPFL ne cesse de croître grâce à l'apport documentaire des chercheurs, collaborateurs et bibliothécaires.

L'archive institutionnelle est divisée en deux volets, d'une part la production scientifique interne et d'autre part le référencement de la documentation des laboratoires.

6.5.2 Production scientifique et Ressources documentaires

Force est de constater que ces deux parties crée la confusion lors d'une première connexion à Infoscience. La partie Ressources documentaires recense des documents hors production EPFL. Elle a été créé pour les bibliothèques de laboratoires dont les ouvrages n'avaient pas été catalogués dans NEBIS. De plus, le volet documentation semble être gelé depuis quelques temps car les laboratoires n'envoient pas systématiquement les références de leurs nouvelles acquisitions. Cette partie d'Infoscience mérite donc d'être repensée et développée afin d'offrir un outil performant pour la recherche de références bibliographiques.

Le débat fait rage au sein de la bibliothèque et du centre informatique au sujet de l'avenir de la partie Ressource documentaire dans son état actuel. Doit-on la conserver en la développant et en l'améliorant ou doit-on tout bonnement la supprimer d'Infoscience ?

L'outil Infoscience est relativement bien pensé en ce qui concerne la gestion documentaire. Il est très structuré et offre des possibilités de recherche par champs appréciables. Le volet Ressource documentaire peut donc être géré de manière convenable sur cette plateforme. Est-il donc réellement nécessaire de supprimer ces ressources ou du moins les déplacer vers un autre système ? La question mérite d'être posée.

Après tous les efforts menés pour promouvoir Infoscience aux yeux de la communauté scientifique, il serait plus judicieux à mon sens de distinguer clairement les deux services sur deux serveurs distincts. Cela aurait l'avantage, tout en gardant le même système informatique, de développer les 2 volets indépendamment l'un de l'autre. Par ailleurs, cela apporterait une meilleure lisibilité et supprimerait toute ambiguïté lors de la consultation.

Une solution subsidiaire consiste en la création d'une gestion électronique des documents (GED) qui à l'avantage, de part la confection d'un plan de classement, de proposer une solution plus souple en terme d'import de types de document. On peut y déposer tout ce que l'on souhaite contrairement à Infoscience qui ne traite pas les revues de presse ou les fichiers vidéo par exemple.

Cependant, la mise en place d'une GED pourrait ne pas trouver grâce aux yeux de la communauté scientifique qui est habituée à travailler avec Infoscience. De plus la gestion de 2 systèmes différents peut accroitre le cahier des charges des informaticiens et bibliothécaires de manière importante.

Dans l'état actuel des choses, il semble plus judicieux de proposer un nouveau service de ressources documentaires basé sur CDS-Invenio. Une séparation claire avec la partie production scientifique EPFL est devenue nécessaire. Afin d'élargir la gamme de documents, il est envisageable d'imaginer de nouveaux formats et types pour ces derniers. Il sera donc nécessaire d'adapter l'outil pour l'import de l'ensemble des ressources documentaires présent sur le site de l'école. Par ailleurs, pourquoi ne pas incorporer le catalogue de l'Université de Lausanne (UNIL) afin d'élargir l'offre documentaire.

Au final, c'est le groupe Infoscience, composé de bibliothécaires et d'informaticiens, qui prendra la décision finale suivant les ressources disponibles au sein de l'école.

6.5.3 Les usagers

Cet outil cible la communauté des chercheurs scientifiques, professeurs et doctorants. Les étudiants Bachelor et Master fréquentent peu le portail d'information car ils ne constituent pas la cible d'Infoscience. Il faut savoir qu'à la base cet outil était destiné aux laboratoires du campus. Les ressources documentaires nécessaires à leur cursus sont stockées sur la plate-forme Moodle. Il faut savoir que dès son origine, Infoscience était dédié aux laboratoires scientifiques. En effet, les étudiants de l'école n'ont en principe pas l'autorisation de déposer leurs travaux sur la plate-forme. Selon Gregory Favre, coordinateur d'Infoscience, le service informatique ne fait pas de publicité pour cette cible d'utilisateurs. Cependant, si un étudiant en fait la demande et que son travail est de qualité, une réflexion pourrait être menée sur ce point. Il faut savoir que les travaux de Master ne représentent pas un réel travail menant à une découverte pour les sections. Ils s'apparentent plus à un examen de fin d'étude. Infoscience fait également office de portail d'information sur le patrimoine de l'EPFL.

Infoscience est principalement consulté lors de l'élaboration de projets de semestre ou de thèses. Cependant, la mise à disposition de documents numérisés consultables en texte intégral permettrait à cet outil de séduire un plus large spectre d'étudiant selon Raphaël Gindrat, président de l'association des étudiants.

6.5.4 Critiques

Lors de mes entretiens au sein de la communauté scientifique, j'ai observé certaines réserves au sujet d'Infoscience. Le fait de centraliser l'information sur une plate-forme unique alimente le spectre de la pérennité des données. Olivier Burdet du laboratoire IBETON conserve sur un disque dur personnel toutes les publications qu'il transmet en vue du dépôt dans Infoscience. Des problèmes de compatibilité des serveurs ont prouvés qu'il était délicat de se fier aveuglement à cette technologie.

Par ailleurs, le module de recherche sur le portail d'information n'est pas optimal, dû principalement à des problèmes de cartographie. Il est vrai qu'Infoscience s'apparente parfois à un supermarché de l'information où l'usager peut se perdre facilement dans la masse d'information.

Actuellement, la recherche plein texte sur un fichier PDF est impossible dans Infoscience. Le module d'indexation automatique proposé par le système CDS-Invenio

n’est pas encore abouti. Il se caractérise par des lenteurs rendant ce processus inopérable. Selon Grégory Favre, le CERN n’utilise pas cette option pour leur gestion documentaire.

Une des remarques que l’on peut faire aux administrateurs d’Infoscience est le manque de communication envers les usagers lorsque des mises à jour importantes sont effectuées. Les laboratoires ne sont pas toujours informés de ces changements.

6.5.5 Mise en ligne des documents

Il est à noter qu’Infoscience a été conçu à l’origine pour diffuser des documents numériques selon Lionel Walter, bibliothécaire scientifique IT. Les fichiers électroniques sont envoyés à Grégory Favre qui est en charge d’alimenter la base de données des publications scientifiques. La structure d’Infoscience nécessite la saisie de nombreux champs bibliographiques (domaine scientifique, nature éditoriale, etc.) pour le dépôt d’un document. Cet outil offre également la possibilité d’ajouter du contenu additionnel tel que des images et des transparents. Le format des documents stockés dans Infoscience est le PDF. Ceci représente le 9/10 des ressources bibliographiques à l’heure actuelle.

Conclusion

Ce travail a pour principal objectif de mener une expertise et de développer un savoir-faire à la bibliothèque en matière d'analyse de fonds à numériser. Après avoir fait un état des lieux des collections étudiées, des pistes de réflexion ont été proposées afin de mettre en lumière les différentes contraintes liées à un tel projet.

La première partie du rapport a pour but de présenter le thème de la numérisation dans son ensemble. Cette technique est maintenant encrée dans notre quotidien et de nombreuses institutions se sont dotées de scanners spécialisés ou ont fait appel à un prestataire de service pour traiter leurs ressources documentaires.

Les contenus et les moyens d'accès à l'information ont fortement évolués au cours de ces dernières décennies. Les besoins et les exigences des consommateurs ont également suivis cette tendance. Les institutions culturelles et scientifiques ont dû s'adapter à ce nouveau paradigme en développant des outils toujours plus perfectionnés.

La numérisation est une solution visant à rendre visible des données difficile d'accès en les regroupant dans des collections virtuelles. Par ailleurs, elle offre de nombreux atouts de recherche tels que la recherche booléenne, la recherche par champ et la recherche dans le texte. La numérisation offre, de surcroît, des opportunités de collaboration entre institutions qui permettront par la suite d'étendre les services aux usagers et d'étoffer la collection.

Nous l'avons vu tout au long du présent rapport, la numérisation répond à de nombreux défis tels que la préservation des données, la perte de données et les capacités insuffisantes de stockage.

La numérisation est un procédé technique faisant appel à des connaissances précises en informatique, en science documentaire et en gestion de projet. La complexité des procédés demande une attention de tous les instants et nécessite un investissement non négligeable en ressources humaines et financières. En outre, tout projet de numérisation implique une connaissance sans faille des collections à traiter.

Ce dernier point ne m'a pas permis d'élaborer des recommandations sur des données tangibles. En effet, la gestion des collections diffère fortement d'un laboratoire à l'autre. L'abandon de certains centres de documentation n'a pas facilité l'identification des collections. De ce fait, il est difficile d'élaborer un plan d'action commun pour chaque

entité. Mon étude se base en grande partie sur des estimations. Le lancement d'un projet de numérisation nécessitera d'entrer plus profondément au cœur des collections tout en collaborant avec les propriétaires de celles-ci.

Mon étude tend à démontrer que la numérisation de masse ne semble pas être adaptée aux collections étudiées. De nombreux documents n'ont pas ou plus de valeur aux yeux des laboratoires et de la bibliothèque de l'EPFL. L'avantage de procéder par sélection de document permettra de faire la lumière sur la valeur réelle des fonds. Certes chronophage, cette méthode a le mérite de « nettoyer » les collections et d'en garder l'essentiel.

Ce rapport souligne également le fait que de nombreux scientifiques connaissent que partiellement le contenu documentaire de leur bibliothèque de laboratoire. Cela se remarque particulièrement au sein des entités ne possédant plus de gestionnaire de collection attitré. En outre, l'abandon de la gestion documentaire signe l'arrêt de mort de ces bibliothèques qui seront probablement remplacées par des dépôts d'archives à terme.

L'une des contraintes les plus marquées dans un projet de numérisation, hors problématique purement technique, est la gestion des droits d'auteurs. Il est difficile de se prononcer clairement sur l'ampleur des documents à traiter sans avoir au préalable consulté les ayants droit. La diffusion de l'information est soumise à la loi sur les droits d'auteur. Ces contraintes légales se présentent comme un frein pour tout projet de numérisation d'envergure et ne facilite en rien le processus.

D'un point de vue personnel, ce travail au sein de la bibliothèque de l'EPFL m'a fait prendre conscience de la complexité de mener à bien un tel projet. La pluralité des acteurs, les collections hétérogènes, les nombreuses contraintes techniques et organisationnelles ainsi que les méthodes de travail spécifiques à chaque entité ne facilitent pas la mise en place d'un tel projet. Le déroulement de cette étude a été une réelle découverte. C'est la première fois que je prends part de manière autonome à une étude d'une telle ampleur. Il n'a pas toujours été aisé d'adapter la théorie à la pratique à cause des nombreuses particularités propre aux organismes étudiés. Cependant, cette expérience m'a tenu en haleine du début à la fin, et ceci malgré les obstacles et autres contraintes temporelles. Il a été, pour moi, impossible de suivre le déroulement initial proposé dans mon cahier des charges. Les nombreux entretiens nécessaires à la compréhension des modes de fonctionnement des différents organismes, a pris énormément de temps. Cependant, je pense avoir répondu aux objectifs fixés dans mon cahier des charges. Il est vrai que je ressens une certaine

frustration de ne pas avoir pu, par manque de temps et par esprit de synthèse, approfondir certaines parties clés de mon étude.

J'espère que les éclaircissements apportés par mon étude auront été utiles à mon mandant et que mes recommandations auront permis de faciliter la prise de décision pour le lancement d'un éventuel projet de numérisation.

Bibliographie

Monographies

BARRELET, Denis. EGLOFF Willi. *Le nouveau droit d'auteur : commentaire de la loi fédérale sur le droit d'auteurs et les droits voisins.* Berne : Ed Stämpfli, 2008. 437 p.

CHAUMIER, Jacques. *Documents et numérisation : enjeux techniques, économiques, culturels et sociaux*. Paris : ABDS, 2006. 119 p. (Sciences et techniques de l'information).

CHEVALIER, Aline. TRICOT, André. *Ergonomie des documents électroniques*. Paris : Presse Universitaire de France, 2008. 305 p. (Le travail humain).

CLAERR, Thierry. WESTEEL, Isabelle. *Numériser et mettre en ligne*. Villeurbanne: Presse de l'ENSSIB, 2010. 200 p. (La boîte à outils).

DUCHEMIN, Pierre-Yves. *L'art d'informatiser une bibliothèque : guide pratique*. Paris : Ed du Cercle de la Librairie, 2000. 587 p. (Bibliothèques).

JACQUESSON, Alain. RIVIER, Alexis. *Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux*. Paris : Cercle de la Librairie, 2005. 573 p. (Bibliothèques).

JACQUESSON, Alain. *Google Livres et le futur des bibliothèques numériques.* Paris : Cercle de la Librairie, 2010. 223 p. (Bibliothèques).

PAPY, Fabrice. *Les bibliothèques numériques*. Paris : Editions Lavoisier, 2005.

VERHEUL, Ingeborg. TAMMARO, Anna Maria. WITT, Steve. *Digital library futures : user perspectives and institutional strategies.* La Haye, IFLA, 2010. 150 p.

Articles scientifiques

AYMONIN, David. RITTMEYER, M. La bibliothèque de l'EPFL au Rolex Learning Center, in : *Bibliothèques d'aujourd'hui ; à la conquête de nouveaux espaces* ; Ouvrage collectif sous la direction de Marie-Françoise Bisbrouck ; Electre 2010, EAN 9782765409823.

ARLITSCH, Kenning. The Espresso Book Machine : a change agent for libraries*, In : Library Hi Tech.* J. Willard Marriott Library, University of Utah, Salt Lake City, Utah, USA, 2011. Vol. 29, No 1, p. 62-72.

DENOREAZ-BUCLIN, Laurence. AYMONIN, David, FAVRE, G., WALTER, L. Enrichir la base des journaux et interface de déduplication, *In : Infoscience*, 2010.

NICHOLSON, Shawn W. PEIFFER, Richard. SHAW, John D. Hardware in libraries : making informed choices*, In : Library Hi Tech.* Michigan State University Library, East lansing, Michigan, USA, 2011. Vol. 29, no 1, p.73-82.

SARDET, Frédéric. *L'offre numérique scientifique en Suisse : questions d'identification.* Arbido, édition 1, 24 février 2011.

Articles de périodique

Histoire du jour IVO IOSSIGER : Son invention fait un carton. C'est une invention suisse révolutionnaire, *In : Le Matin*. 29/07/2003.

Flash informatique n° 1/2009 et 7/2010. (Périodique EPFL).

Comptes-rendus de conférence

AYMONIN, David. *La Bibliothèque de l'EPFL au Rolex Learning Center : 10 mois après l'ouverture. Journées d'étude : "Learning centres : vers un modèle à la française ?"*, MEDIAT Rhones Alpes, Univeristé Claude Bernard, Lyon, France, December 6-7, 2010.

AYMONIN, David. GUIGNARD, T. *La bibliothèque de l'EPFL au Rolex Learning Center enfin révélée (Revealed at last).* Invitation des collègues bibliothécaires de Suisse et d'Europe dans le cadre de l'inauguration du RLC, Lausanne, Suisse, May 28, 2010.

AYMONIN, David. *La politique ou l'utilité ? Quelques observations sur les facteurs d'adoption des archives institutionnelles par les chercheurs et leurs universités et leurs conséquences sur le travail des informaticiens et des bibliothécaires. Penser global, agir local.* Politiques de mise en ligne de la production académique, Université de Nice Sophia Antipolis, 29-30 Mars 2010.

Document numérique et société : actes de la conférence DocSoc – 2006, Semaine du document numérique / Sous la dir. De Ghislaine Chartron et Evelyne Broudoux. Paris : Association des professionnels de l'information et de la documentation (ADBS), 2006. 342 p.

SIMIONI, O. AYMONIN, David. *Bibliothèque et utopie «Comment classer le monde…»*. Manifestation culturelle BCU-Bibliothèque de l'EPFL, Rolex Learning Center, EPFL, Lausanne, December 8, 2010.

The DSpace Open Source Digital Asset Management System: Challenges and Opportunities

Rapports d'expertise

Final Report: Comparison between CDSWare and DSpace, with some specific conclusions concerning CDSWare. Rapport effectué par FontisMedia.

Webographie

4DIFITAL BOOKS. *4DigitalBooks* (en ligne). 2011. http://www.4digitalbooks.com/ (consulté le 14.04.2011).

BIBLIOTHEQUE NATIONALE DE FRANCE. *Bibliothèque nationale de France* (en ligne). 2011. http://www.bnf.fr/fr/acc/x.accueil.html (consulté le 22.05.2011).

BIBLIOTHEQUE DE L'UNIVERSITE DE CORNELL. *De la théorie à la pratique : didacticiel d'imagerie numérique* (en ligne). 2000. 2003.

http://www.library.cornell.edu/preservation/tutorial-french/contents.html (conuslté le 15.07.2011).

DIGICOORD. *DIGICOORD* (en ligne). https://www.digicoord.ch/index.php/Accueil (consulté le 25.09.2011).

E-LIB.CH. *e-lib.ch* (en ligne). 2011. *http://www.e-lib.ch/fr/* (consulté le 25.09.2011).

ENSSIB. *ENSSIB* (en ligne). 2011. http://www.enssib.fr/ (consulté le 05.06.2011).

ENSSIB. *Eléments pour l'appréciation des coûts : étude réalisée par Benoit Epron* (en ligne). 2002. http://pisrsh.mmsh.univ-aix.fr/couts.htm (consulté le 15.07.2011).

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS. *IFLA* (en ligne). 2011. http://www.ifla.org/ (consulté le 05.06.2011).

MINISTERE DE LA CULTURE. *Informations techniques* (en ligne). http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_04.htm (consulté le 15.07.2011).

Plume : promouvoir les logiciels utiles maîtrisés et économiques dans l'enseignement supérieur et la recherche (en ligne). 2010. http://www.projet-plume.org/ (consulté le 15.07.2010).

SECURARCHIV SA. *Secur'Archiv : le gestionnaire de votre information* (en ligne). 2006.2011. http://www.securarchiv.ch/index.php (consulté le 15.05.2011).

Annexe 1
Glossaire

IREC (1971-2001)

L'Institut de recherche sur l'environnemement construit rassemble une équipe interdisciplinaire de chercheurs qui analyse l'environnement construit et le phénomène urbain sur les plans régional, national et mondial, mais aussi à l'échelle des quartiers et des bâtiments. L'institut contribue ainsi à une meilleure connaissance de ces structures et processus économiques, sociaux, politiques et culturels, et à l'élaboration de stratégies d'action. A travers ses cours et ses publications, l'IREC sensibilise les architectes et ingénieurs – tant étudiants que professionnels –, l'opinion publique et les milieux politiques aux transformations et futurs enjeux spatiaux. Il contribue à une analyse véritablement interdisciplinaire grâce à son insertion dans l'EPFL et à ses échanges directs avec les bâtisseurs de la ville et du territoire.

CEAT

Anciennement IREC, la CEAT est la communauté d'étude de l'aménagement du territoire. La CEAT est une plateforme de coordination pour la recherche au niveau romand, rattachée à la faculté de l'environnement naturel, architectural et construit (ENAC). La Communauté d'étude a été créée par les cantons romands en 1970 dans le but de regrouper les professionnels de l'aménagement du territoire. Cette institution scientifique travail sur des mandats émanant des Villes, des Cantons et de la Confédération dans le domaine de la recherche. En parallèle à cela, la CEAT est active dans l'enseignement et la recherche.

CEDEC

Le Centre de Documentation de l'Environnement Construit.

ENAC

Faculté de l'environnement naturel, architectural et construit. Elle se compose de:

- 3 sections d'enseignement
- 4 instituts de recherche [Institut d'architecture et de la ville (IIC), Institut d'ingénierie civile (IIC), Institut de l'urbain et du territoire (INTER), Institut d'ingénierie de l'environnement (IIE)
- 1944 étudiants, dont 242 doctorants contribuant à la recherche en ENAC
- 619 collaborateurs, ou 488 équivalents plein-temps
- 74 millions CHF, budget total de l'ENAC

ICOM

L'Institut de la construction métallique est une unité faisant partie de l'Institut d'ingénierie Civile (IIC). La mission de l'institut a été définie de la manière suivante: enseigner la conception et le dimensionnement des structures métalliques, mener une recherche théorique et appliquée dans ce secteur et proposer ses services dans le cadre de travaux pour tiers.

IBETON

Au sein de la faculté ENAC, section de Génie Civil de l'Ecole Polytechnique Fédérale de Lausanne, l'équipe du Laboratoire de construction en béton est active dans le domaine des structures en béton. Ses missions principales sont l'enseignement, la recherche et le service au tiers (expertises, mesures...).

LMH

Le laboratoire des Machines Hydrauliques de l'EPFL intervient aux quatre coins de la planète comme expert indépendant pour vérifier que le rendement des turbines soit optimal. Les missions principales du laboratoire sont l'enseignement, la recherche et le service aux tiers dans le domaine de l'hydrodynamique.

Les images numériques

Ces spécifications techniques se basent sur le didacticiel d'imagerie numérique de la bibliothèque de Cornell ainsi que sur l'ouvrage de Thierry Claerr et Isabelle Westeele, « numériser et mettre en ligne ».

Les images numériques sont des clichés électroniques d'une scène ou numérisés à partir de documents tels que photographies, manuscrits, textes, imprimés, et œuvres d'art. Cela concerne toutes les images acquises, créées, traitées et stockées sous forme binaire. L'image numérique est échantillonnée et mappée comme une grille de points ou éléments d'images. A chaque pixel correspond une valeur tonale (noir, blanc, niveaux de gris ou couleur), exprimée en code binaire (0 et 1). Les chiffres binaires (bits) de chaque pixel sont stockés dans une séquence par l'ordinateur et souvent réduit à une expression mathématique (compressée). Les bits sont alors réinterprétés et lus par l'ordinateur afin de délivrer une version analogique en vue d'être affichée ou imprimée.

La résolution

La résolution est la capacité à distinguer les détails fins dans l'espace. C'est une mesure de la finesse de l'affichage ou de la capture d'une image, exprimée en nombre de pixels par unité de surface, c'est à dire la « densité » en pixels. La fréquence spatiale à laquelle une image numérique est échantillonnée (fréquence d'échantillonnage) est généralement un bon indicateur de la résolution. C'est pourquoi les termes points par pouce ou pixels par pouce (ppi ou dpi) sont les expressions courantes et synonymes indiquant la résolution des images numériques. En général, l'augmentation de la fréquence d'échantillonnage augmente la résolution, mais seulement jusqu'à un certain point.

Dimensions en pixel

Ce sont les mesures horizontales et verticales d'une image exprimées en pixel. Les dimensions en pixel peuvent être déterminées en multipliant la largeur et la longueur de l'image par le dpi. Un appareil photo numérique possède également des dimensions en pixel. Le nombre de pixel horizontaux et verticaux définissent sa résolution. Pour calculer la résolution en dpi, il faut diviser une des dimensions en pixel par la dimension en pouces correspondante.

Un document de 8 x 10 numérisé à 300 dpi a une résolution en pixels de 2400 pixels (8x300 dpi) par 3000 pixels (10X300 dpi).

La résolution sera donc choisie selon la taille de l'original, l'utilisation prévue, les capacités du système d'archivage et ses évolutions.

Voici un tableau récapitulatif des résolutions pour une conservation à long terme :

support	cas général	exception	remarques
Originaux opaques au format A6 ou supp. (imprimés, manuscrits, calques...)	300 ou 400 dpi	600 dpi	un scan à 600 dpi est utile si le doc. présente des variations de formats ou d'informations
Originaux opaques au format inférieur à A6 (cartes postales, médailles et monaies...)	600 dpi		
Reproductions transparentes (ektas, diapositives, microformes...)	300 ou 400 dpi	600 dpi	

La profondeur en pixel (ou profondeur d'acquisition ou profondeur de couleur)

La profondeur est définie par le nombre de bits utilisés pour représenter chaque pixel. Plus la profondeur de bit est élevée, plus grand sera le nombre de teintes (niveaux de gris ou couleur) représenté. Les images numériques peuvent être produites en noir et blanc (deux couleurs), niveaux de gris ou couleur.

Une *image bitonale* (deux couleurs) est représentée par des pixels de 1 bit chacun, pouvant représenter deux teintes (d'habitude le noir et le blanc), en utilisant la valeur 0 pour le noir et 1 pour le blanc.

Une *image en niveaux de gris* est composée de pixels possédant plusieurs bits d'informations, allant en général de 2 à 8 bits, ou davantage.

Une *image couleur* est typiquement représentée par une profondeur de bit variant de 8 à 24 bits ou plus. Dans une image 24 bits, les bits sont souvent divisés en 3 groupes : 8 pour le rouge, 8 pour le vert et 8 pour le bleu. Les combinaisons de ces bits servent à représenter les autres couleurs. Une image 24 bits offre 16,7 millions de valeurs de couleurs (2^{24}). De plus en plus, les scanners capturent chaque canal de couleur à 10 bits ou plus, et les réduisent à 8 bits afin de compenser le "bruit" du scanner et présenter une image aussi proche que possible de la perception visuelle de l'être humain.

Calculs binaires pour le nombre de teintes représentées par les profondeurs de bit courantes :

1 bit ($2^{1)}$ = 2 tons

2 bits (2^2) = 4 tons

3 bits ($2^{3)}$ = 8 tons

4 bits ($2^{4)}$ = 16 tones

Etc.

La profondeur d'acquisition est donc fonction du support et de son contenu. Par exemple, le niveau de gris sera adapté à la presse car les caractères sont très fins et la qualité du papier hétérogène.

Profondeur d'acquisition	support
Couleurs	Tout ou majoritairement en couleur
Niveaux de gris	Documents en demi-teintes (lavis, fusain, photographies, dessins à détails très fins. Imprimés contenant un grand nombre de photographies : presse par exemple. Documents avec de fortes rousseurs, très tâchés sur la zone imprimée ; faiblement contrastés ; à l'impression irrégulière.
Noir et blanc	Imprimés courants : schémas, tableaux ; dessins au trait.

La taille de fichier

La taille de fichier est calculée en multipliant la surface d'un document (hauteur x largeur) à numériser par la profondeur de bit et le dpi au carré. Parce que la taille d'un fichier image est exprimée en bytes, qui sont composés de 8 bits, divisez ce chiffre par 8.

Si les dimensions en pixels sont données, multipliez-les entre elles et par la profondeur de bits pour définir le nombre de bits d'une image. Par exemple, si une image 24 bits est capturée avec un appareil photo numérique aux dimensions de 2048 pat 3072, alors la taille du fichier est égale à (2048 x 3072 x 24)/8, soit 18.874.368 bytes.

Convention sur le Nommage de la Taille de fichier : parce que les images numériques résultent fréquemment en des fichiers très larges, le nombre de bytes est généralement représenté par incréments de 2^{10} (1024) ou plus.

1 Kilo-octet (KB ou Ko) = 1024 bytes ou octets

1 Megaoctet (MB ou Mo) = 1024 KB (ou Ko)

1 Gigaoctet (GB ou Go) = 1024 MB

1 Teraoctet (TB ou To) = 1024 GB

Format et Compression

Un format de fichier est une manière de coder les données. Il peut être propriétaire ou non. Certains formats sont plus adaptés à la numérisation que d'autres. Le format RAW (brut) est un format natif propre au constructeur généré par les appareils photos. C'est une sorte de négatif de l'image. Plusieurs informations sont contenues dans ce format tel que les conditions de prise de vue et les différents réglages. Il permet d'évaluer la qualité d'une image avant son traitement mais n'est pas un format d'archivage. Son utilité est donc limitée.

Il existe une multitude de format, cependant la référence en la matière reste le format TIFF (**T**ag **I**mage **F**ile **F**ormat), très utilisé pour l'archivage à long terme. Cependant, il est peu adapté à la diffusion car les fichiers générés sont très lourd et certains navigateurs ne le supportent pas sous sa forme native. On lui préférera le format PDF pour la diffusion et l'échange de fichiers sur Internet. Par ailleurs, ce format propose des fonctions de recherche plein texte pour les projets incluant l'OCR.

La compression est utilisée afin de réduire la taille de l'image pour le stockage, le traitement et le transfert. La taille des fichiers d'images numériques peut être assez conséquente, ralentissant les capacités informatiques et réseau de nombreux systèmes. Tous les schémas de compression réduisent la chaîne de code binaire de l'image non compressée en une formule mathématique raccourcie basée sur les algorithmes. Il existe des schémas de compression standard et d'autres propriétaires. En général il est préférable d'utiliser un schéma standard largement supporté qu'un schéma propriétaire qui bien qu'offrant un schéma de compression plus évolué et/ou d'une meilleure qualité, pourrait ne pas se prêter à une utilisation à long terme ou aux stratégies de préservation numérique.

Les schémas de compression peuvent, de plus, être décrits comme destructifs ou non destructifs. Les schémas non destructifs, tels que le ITU-T.6, réduisent le code binaire sans perdre d'informations, ainsi lorsque l'image est décompressée, elle est identique à l'original, bit pour bit. La compression destructive, telle que le JPG, utilise des moyens afin de niveler ou écarter les informations les moins significatives, sur la base de la perception visuelle de l'être humain. Néanmoins, il peut être extrêmement difficile de détecter les effets de la compression destructive, et l'image peut être considérée comme "visuellement intacte". La compression non destructive est généralement plus souvent utilisée avec les numérisations bitonales de documents textuels. La compression destructive est généralement utilisée avec les images tonales, et en particulier les images à tons continus où la réduction d'informations ne résulte pas en un allègement appréciable du fichier.

Si l'image est destinée à être réutilisée et archivée, il est préférable de choisir la compression sans perte consistant généralement à factoriser les pixels identiques et à mémoriser leur emplacement dans l'image pour les restituer à leur place à la décompression. Depuis peu, un nouvel algorithme de compression à vu le jour. Il s'agit du JPEG 2000 avec ou sans perte qui est plus puissant que le JPEG traditionnel. Il permet la création d'images de résolutions et de tailles différentes ce qui l'apparente aux formats de diffusions. De plus grâce à sa compression sans perte, il est parfaitement adapté à un archivage à long terme et il réduit fortement les espaces de stockage nécessaires à l'archivage.

Annexe 2
Tableau récapitulatif des principaux formats de fichiers

Nom / producteur	Documentation	caractéristiques	Compression	Gestion des couleurs	Utilisation préconisée
TIFF – Tagged Image File Format / Aldus Corp. (rachat par Adobe) et Microsoft	ISO 12639 : 2004. Spécifications (v. 6, 1992)	Extensions : .tiff, .tif Monopage (un fichier, une image) ou multipage (un fichier pour une séquence d'images) Archivages d'images lourdes indépendamment des plateformes et des périphériques. Largement supporté par les applications de gestion d'images	LZW UIT Groupe 4 (réversible) Chaque format compressé selon un algorithme forme un sous-ensemble du TIFF v.6.	Etendue (jusqu'à 24 bits). Supporte plusieurs espaces de couleur et plusieurs profils ICC	Très utilisé pour l'archivages à long terme (souplesse d'utilisation, bonne gestion des couleurs, documentation). Peu adapté pour la diffusion (lourdeurs des fichiers). Tous les navigateurs ne le supportent pas encore sous sa forme native. Compression UIT Groupe 4 seulement pour les images bitonales
JFIF – JPEG File Interchange Format / C-Cube Microsystems	JFIF JPEG : ISO : IS 10918-1 – Recommandation T.81	Extension : .jpeg, .jpg, .jpe, .jfif, .jfi Format d'image pour les fichiers compressés avec l'algorithme JPEG (compression avec perte pour échanger des images sur les réseaux avec une qualité acceptable)	JPEG. Taux à choisir : 100% = pas de compression ; 75-80% fournit un bon ratio	Etendue (jusqu'à 24 bits). Utilisé de préférence pour les photos	Peu adapté pour l'archivage. Conçu pour la diffusion et l'échange d'images sur Internet, entre de multiples plateformes et applications
JPEG 2000 / Joint Photographic Experts	ISO – Technologies de l'information – SC 29 – UIT	Extension : .jp2 Ensemble des méthodes de compression	JPEG 2000	Très étendue (jusqu'à 32 bits).	Utilisation accrue à partir de 2005. Utile

Group, ISO, UIT		avec ou sans perte et de format de fichier		Supporte plusieurs espaces de couleurs et plusieurs profils ICC	pour la diffusion car génère des image de résolution variables, ce qui évite la gestion d'un jeu de fichiers de diffusion
PDF – Portable Document Format / Adobe	ISO 32000-1 :2008 (v 1.7)	Extension : .pdf Créé d'abord pour l'impression. Indépendant du système d'exploitation. Peu englober plusieurs contenus dans d'autres formats et comporter des formulaires.	Oui, plusieurs algorithmes	Etendue Supporte plusieurs espaces de couleur et plusieurs profils ICC	Adapté pour la diffusion et l'échange de fichier sur Internet, textuel surtout ; peu adapté pour les images lourdes. Fonctions de recherche en plein texte utile pur les projets incluant l'OCR.

Source : CLAERR, Thierry, WESTEEL, Isabelle. Numériser en mettre en ligne, p. 28-29

Annexe 3
Photographies : 4DigitalBooks


Scanner feuille à feuille.


Scanner à balayage horizontal.


Traitement du fichier image sur le poste informatique.


Scanner automatique haute performance.

Traitement d'anciens périodiques.

 Atelier de production.

Annexe 4
Photographies : Atelier de numérisation de la Ville de Lausanne

Scanner Heidelberg.


Scanner à plans.


Plotter.

 Studio de photographie.

Annexe 5
Photographies : Collection de tirés-à-parts de mathématique


Compactus.


Contenu d'une boîte d'archive (thèses et articles de périodiques).

Annexe 6
Photographies : Documentation CEAT


Armoires de rangement situées dans les couloirs.


Compactus situés dans le dépôt d'archive.

Annexe 7
Photographies : Bibliothèque IBETON

Etagères principales de la bibliothèque.

Ce cliché présente la collection du Beton Kalender édité par Ernst & Sohn dont la série prend fin en 1993 au sein de la bibliothèque. Cette collection présente la vision de l'ingénieur sur les structures et constructions en béton. Ces ouvrages ne sont plus consultés au laboratoire.

Annexe 8
Photographies : Bibliothèque ICOM

Etat des lieux de la bibliothèque.

Poste de consultation.

Bureau du bibliothécaire.

Thèses et autres publications.

Annexe 9
Photographies : Bibliothèque du LMH

Etat des lieux de la bibliothèque

Compactus et présentoir à périodiques.

Annexe 10
Offre de prestation de l'entreprise 4DigitalBooks suite à mon appel d'offre. Extrait de la correspondance avec M. Rod

« Je désire recevoir une estimation des tarifs pour la prise en charge de divers lots de documents. Dans le cadre de mon étude, je dois traiter :

- Une collection de tirés à part de mathématique unique (6000 documents au format A4 et A5)
Les pièces les plus anciennes datent de 1910 (cahiers, feuillets). Le papier a jauni et les reliures se désagrègent. Besoins : Niveaux de gris, 300 dpi, TIFF/PDF.

Scanning : 0.30 par page

Séparation des pages et détourage 0.05 par page

Traitement d'image : 0.05 par page

(OCR (si nécessaires) 0.10 la page

Agrégation et nommage fichiers 1.5o par document

Prise en charge par document (livre, classeur, boîte d'archives) 5.00

- Une collection de monographies (750 documents - 200 p. de moyenne par doc - au format A4)
Besoins : Gestion des couleurs, 300 dpi, PDF.

Scanning 0.50 page

Séparation des pages et détourage 0.05 par page

Traitement d'image : 0.05 par page

(OCR (si nécessaires) 0.10 la page

Agrégation et nommage fichiers 1.5o par document

Prise en charge par document (livre, classeur, boîte d'archives) 5.00

- Un lot de publications scientifiques (250 documents au format A4) Besoins : Gestion des couleurs, 300 dpi, PDF.

Scanning 0.50 page

Séparation des pages et détourage 0.05 par page

Traitement d'image : 0.05 par page

(OCR (si nécessaires) 0.10 la page

Agrégation et nommage fichiers 1.5o par document

Prise en charge par document (livre, classeur, boîte d'archives) 5.00

- Un lot de plans directeurs (60 documents au format A0 et A1) Besoins : Gestion des couleurs, 300 dpi, PDF.

Scanning A1 : 10.oo par doc.

Scanning A0 : 20.00 par doc.

Détourage : 2.00 par doc

Traitement d'image : 2.00 par doc.

Nommage des fichiers

Prise en charge par lot de documents :

- Un lot de photographies sous forme de diapositives (100 pièces)

Nous ne faisons pas de scanning de diapositives. Je ne peux pas vous donner de tarif.

- Un lot de feuillets (100 documents au format A5) Besoins : Niveaux de gris, 300 dpi, PDF.

Scanning : 0.20 par page

Détourage : 0.05 par page

Traitement d'image : 0.05 par page

(OCR (si nécessaires) 0.10 la page

Agrégation et nommage fichiers 1.5o par document

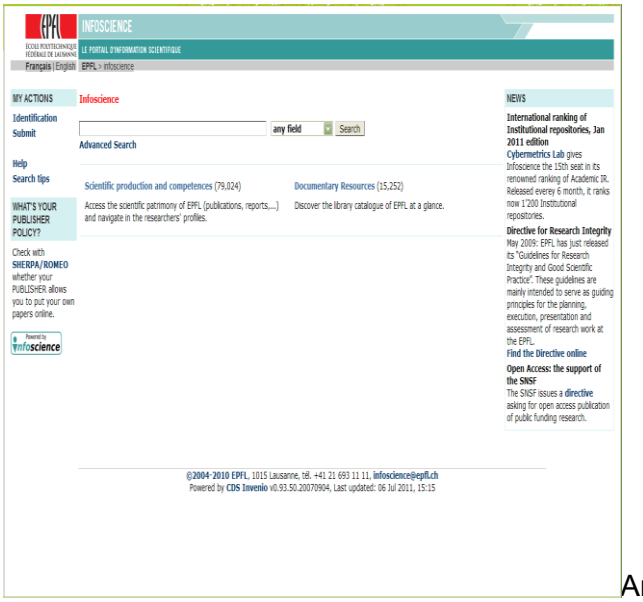
Prise en charge par document (livre, classeur, boîte d'archives) 5.00

Les prix sont en CHF. Livraison des données : 300 CHF pour l'archivage et le groupage sur disque dur externe. Ce poste tombe si vous êtes en mesure de nous fournir un disque dur à remplir par nos soins. Frais de transport : si les documents sont à l'EPFL, je pense que nous pouvons les chercher et les ramener sans frais. Je crois par ailleurs que l'EPFL dispose d'un véhicule de livraison.

Délai de livraison : Pour un projet de cette grandeur (je ne connais pas le nombre de pages du 1er lot, vous parlez de documents mais sans préciser le nombre de pages) il faut compter deux mois.

Le contrôle de qualité est compris dans ces prix. La garantie de qualité à 100% s'applique. »

Annexe 11
Infoscience



Ancienne version d'Infoscience.

Nouvelle version d'Infoscience. Le design est en accord avec la nouvelle charte graphique du site web.